A Latent Trait Look at Pretest-Posttest Validation of Criterion-referenced Test Items

Wim J. van der Linden

Twente University of Technology, The Netherlands

Since Cox and Vargas (1966) introduced their pretest-posttest validity index for criterion-referenced test items, a great number of additions and modifications have followed. All are based on the idea of gain scoring; that is, they are computed from the differences between proportions of pretest and posttest item responses. Although the method is simple and generally considered as the prototype of criterion-referenced item analysis, it has many and serious disadvantages. Some of these go back to the fact that it leads to indices based on a dual test administration- and population-dependent item p values. Others have to do with the global information about the discriminating power that these indices provide, the implicit weighting they suppose, and the meaningless maximization of posttest scores they lead to. Analyzing the pretest-posttest method from a latent trait point of view, it is proposed to replace indices like Cox and Vargas' D_{pp} by an evaluation of the item information function for the mastery score. An empirical study was conducted to compare the differences in item selection between both methods.

As in any other area of educational and psychological measurement, more attention has been paid to reliability than to validity aspects of criterion-referenced measurement. Several test parameters have been proposed and compared with their norm-referenced counterparts, assessment methods have been introduced and examined using both real and simulated data, and the criterion-referenced reliability problem seems on its way to a great diversity of solutions (Hambleton & Novick, 1973; Huynh, 1976a, 1976b; Livingston, 1972; Marshall, 1975; Mellenbergh & van der Linden, 1979; Subkoviak, 1976, 1978; Swaminathan, Hambleton, & Algina, 1974; van der Linden & Mellenbergh, 1978; Wilcox, 1978).

That less powerful efforts have been made to tackle the validity problem may in part be due to a standpoint advocated by, for example, Millman (1974). According to this standpoint, criterion-referenced validity is the same as content validity and to establish this the construction of a well-defined domain of items is sufficient. Once an item is included in the domain no empirical information or item analysis can or

Thanks are due to Fred N. Kerlinger, Gideon J. Mellenbergh, Robert F. van Naerssen, and Egbert Warries for their helpful comments; to Hans van Aalst, Fred Boesenkool, Kees Hellingman, Ton Heuvelman, Rien Steen, Niels Veldhuizen, Ronny Wierstra, and Theo Wubbels for participating in the empirical study and computational assistance; and to Paula Achterberg for typing the manuscript.

should lead to reformulating the item or eliminating it from the domain. This *a priori* approach (Popham, 1971) can mostly be found in connection with the idea that item domains are best constructed using item forms (Hively, 1974; Hively et al. 1973; Hively, Patterson, & Page, 1968; Osburn, 1968). Others (e.g., Henrysson & Wedman [1973]; Wedman [1974a]; Smith [Note 1]) take a more empirical view and argue that item writing nearly always involves some subjectivity and that, to eliminate the consequences of this, analysis of empirical item data is indispensable. In this paper we share this empirical approach.

Several attempts have been made to solve the validity problem and to define parameters that can be estimated to ascertain the validity of criterion-referenced items. One contribution is an extensive class of coefficients based on a preinstruction-postinstruction method of test administration. The interpretation, criticism, and replacement of these coefficients are central concerns of this paper. Other contributions are given in Berk (1980), and in several reviews and studies by Hambleton and his colleagues (Hambleton & Gorth, 1971; Hambleton, Swaminathan, Algina & Coulson, 1978; Rovinelli & Hambleton, 1973, 1977; Coulson & Hambleton, Note 2). These authors distinguish three approaches to the validity problem. The first is the aforementioned item form or item generation rule approach. In it, the fixed syntactical structure and variable elements of item sentences are used to define domains and—eventually with the aid of the computer—to sample items. Item validity is automatically guaranteed because the definition of the domain and the construction of the items are accomplished by the same set of rules. The second approach is a judgmental procedure in which content specialists are retained. The judgmental task may assume a form consisting of experts directly rating the item-objective congruence, but semantic differential techniques and matching procedures can also be used. Rovinelli and Hambleton (1977) give an index of item-objective congruence to be used in the direct rating technique and report a comparative study of these three judgment methods. The final approach is the empirical approach. The class of pretest-posttest validation procedures is the best known representative of this approach. Other possibilities are the use of standard (norm-referenced) item statistics, Brennan and Stolurow's suggestion to use Cochran's *Q*, Guttman's scalogram analysis, and factor analysis (see Berk, 1980, part two; Rovinelli & Hambleton, 1973).

As already mentioned, in this paper the validity coefficients based on a pretest-posttest administration are central. Cox and Vargas (1966) were the first to propose a pretest-posttest coefficient for criterion-referenced item validity; additions and modifications have followed. All these methods suppose gain scoring; that is, they lead to coefficients computed on differences between posttest and pretest scores. The pretest-posttest method is simple and often said to be the prototype of criterion-referenced item analysis. Having discussed its rationale, some of the coefficients based on it and a few empirical studies in the following section, we will, however, show that it has serious drawbacks and causes difficulties in interpreting the results. Thereafter, we present an analysis of the pretest-posttest method adopting a latent trait view and propose the use of an item parameter that does not have the aforementioned drawbacks and interpretation difficulties. Finally, using item data collected in an empirical study, we compare the differences in item selection between the pretest-posttest and latent trait approach to the criterion-referenced validity problem.

Pretest-Posttest Method

In traditional item analysis one is generally concerned with statistics like (classical) item difficulty and discriminating power (Lord & Novick, 1968, chap. 15). The former is the item mean score for a group of examinees, the latter the point biserial item-test correlation. When item analysis has to be performed by hand, however, the computational work involved in item-test correlations is mostly circumvented by computing the statistic D . This is the difference in item mean score between the upper and lower 27 percent of the test score distribution. Though it is usual to interpret D as a measure of association between item and test score, Findley (1956) provides a second interpretation of D as the net proportion of correct discriminations between the upper and lower groups. Cureton (1957) shows that the 27 percent, which stems from an assumption of normality, must be replaced, for example, by a 33 percent rule in the case of a uniform test score distribution. When no guessing is involved, it is a classical item analysis rule to require all items to have a difficulty of .50 and as large as possible a value for their discriminating power parameter. It is well known that this rule leads to test maximally differentiating among examinees and score distributions with large variance, and it is precisely this property which makes traditional item analysis well suited for norm-referenced measurement (Carver, 1970; Popham & Husek, 1969). For the sequel it is important to observe that traditional item analysis maximizes the correlation between item and total test scores, and thus, in fact, uses a criterion inherent in the score distribution. In particular, no fixed point on the measurement scale is used, nor is there an external criterion.

The D_{pp} Coefficient

In criterion-referenced measurement one is primarily concerned with the degree to which an examinee has mastered certain specific behaviors—or, to use Glaser and Nitko's (1971) often cited definition, "in measurements that are directly interpretable in terms of specified performance standards" (p. 653). Mere discrimination between examinees, without any other reference than to the mean score of a norm group, is less important. Because in education the specification of performance standards is mostly in the form of learning objectives, Cox and Vargas (1966) proposed a criterion-referenced counterpart to the statistic D to measure the quality of items, not by their sensitivity to test score differences between high and low groups, but to objective-based instruction. Table I displays the bivariate distribution of item responses obtained at a test administration to the same students before and after an instructional unit. The statistic Cox and Vargas propose is the difference between pretest and posttest p values:

$$D_{pp} \equiv p_{.1} - p_{1.}, \quad (1)$$

and this reflects the sensitivity of the item to instruction. Carver (1974) calls items that show an increase in p value from a pretest to a posttest edumetric as distinct from psychometric items which, for the case of no guessing, have a posttest p value of .50. In the literature D_{pp} is called both a *discrimination* index and a *validity* index. This is correct, since it relates differences in item responses to an external criterion, namely the (objective-based) instruction which has taken place in between. A discussion of the rationale underlying Cox and Vargas' proposal can be found in Cox (1971), Ozenne (1971), Edmonston and Randall (1972), Henrysson and Wedman (1973),

TABLE I
Distribution of Item Responses at Pretest-Posttest Administration

Pretest	Posttest		
	$k = 0$	$k = 1$	
$j = 1$	p_{10}	p_{11}	$p_{1.}$
$j = 0$	p_{00}	p_{01}	$p_{0.}$
	$p_{.0}$	$p_{.1}$	

Roudabush (1973), Wedman (1973, 1974a, 1974b), and in the papers by Hambleton and colleagues referred to in the preceding section. Nicolich, Nicolich, and Raph (1974) have proposed to base criterion-referenced item analysis on an item change score. For each item and examinee, this score is defined to be +1 if the item is answered incorrectly on the pretest and correctly on the posttest, -1 if the reverse holds, and 0 otherwise. The index they propose is the average item change score for a population of examinees. However, although these authors do not realize it, this average is easily seen to be equal to Cox and Vargas' D_{pp} .

A Few Modifications

The literature after Cox and Vargas (1966) has yielded a great many variants and modifications of index (1) all somehow based on pretest-posttest data or differences between proportions and presented as criterion-referenced validity coefficients. As is clear from definition (1), the possible increase in p value is in part dependent on the level of mastery the examinees already show at the pretest. A high pretest p value allows only a low D_{pp} value, and, conversely, given a reported low D_{pp} value, it is uncertain whether this results from a high pretest performance or an item possessing hardly any sensitivity to a supposed learning process. For this reason, Brennan and Stolurow (1971) replace D_{pp} by a coefficient indicating the percentage of maximum possible gain:

$$\frac{p_{.1} - p_{1.}}{1 - p_{1.}} \quad (2)$$

They did not realize, however, that McGuigan and Peters (1965) earlier proposed the same coefficient for use with assessing the effectiveness of programmed texts. Assuming a knowledge or random guessing model, Roudabush (1973) derives

$$\frac{(p_{01} - p_{10})p_{0.}}{p_{.0}^2 + (p_{01} - p_{10})p_{0.}} \quad (3)$$

as the proportion of examinees having the item wrong at the pretest and right at the posttest after a correction for guessing has been applied. Kosecoff and Klein (1974) introduce two item validity coefficients, one being the difference between the proportion of examinees who pass the item on the posttest and fail it on the pretest and the proportion who fail it on both occasions,

$$p_{01} - p_{00} \quad (4)$$

The other is a coefficient with the same structure but with proportions of masters in terms of total test scores and computed only from the group of examinees who pass the item at the posttest. The idea of pretest-posttest validation of test items has also been modeled using analysis of variance with repeated measures from both item administrations (Herbig, 1975, 1976). The natural coefficient belonging to this approach is the proportion of total item variance explained by the repeated measures. It can be shown to assume the form

$$\frac{(p_{01} - p_{10})^2}{(p_{0.} + p_{.0})(p_{1.} + p_{.1})} \quad (5)$$

for $p_{01} > p_{10}$, and is equal to zero otherwise (Herbig, 1975). Herbig shows the same concern as Brennan and Stolurow (1971) and also presents a version of (5) free from pretest knowledge. This is obtained by putting $p_{1.}$ equal to zero and replacing $p_{.1}$ by $p_{01} - p_{10}$, which leads to

$$\frac{p_{01} - p_{10}}{p_{0.} + p_{.0}} \quad (6)$$

(Herbig, 1975, 1976). Table I contains only examinees producing a (right or wrong) response on both item administrations. In contrast, Harris (1976) also considers examinees who have not attempted the item and defines the ratio of the numbers of examinees showing a gain score to the number not attempting the item on the pretest as a validity index. It is based on the (unrealistic) assumptions that there is no guessing at the pretest and that examinees who are not able to produce the right answer do not respond. Moreover, unlike the preceding coefficients, Harris' index can assume values outside the intervals $[0, 1]$ or $[-1, +1]$ which the preceding coefficients are restricted to. Instead of using a pretest-posttest design, the simultaneous use of separate groups of experts and nonexperts is sometimes used (see, e.g., Henrysson & Wedman, 1973). Millman (1974) refers to the former as preinstruction-postinstruction (the same individual tested twice) and the latter as instructed-uninstructed criterion groups (different individuals in both groups). With criterion groups of different sizes Cox and Vargas' D_{pp} amounts to an application of Brennan's discrimination index B, which has been shown to be similar to Pierce's theta (Brennan, 1972; Harris & Wilcox, 1980). The possibility of using McNemar's chi-squared

$$\frac{(p_{01} - p_{10})^2}{p_{01} + p_{10}} \quad (7)$$

as an item validity index has been considered by Herbig (1976). Popham (1971) offers a chi square which can be used for determining whether the pattern of proportions in Table I deviates from the average pattern for a domain of items in which the concern is to detect atypical rather than sensitive items. As already mentioned, D_{pp} is widely considered the criterion-referenced counterpart of the traditional statistic D . Sustaining this analogy, Saupe's correlation between item change and total change score should be called the counterpart of the traditional item-test correlation (Saupe, 1966). Although originally presented for the more general problem of selecting item to measure change, it is called so in the criterion-

referenced literature. There are also validity coefficients and procedures based on three item administrations, one prior to the instruction and two afterwards (Brennan, 1974; Brennan & Stolurow, 1971; Ivens, 1972), and, analogous to the pretest-posttest method for item validation, Ozenne (1971) and Barta, Ahn, and Gastright (1976) propose coefficients and procedures for validating tests.

Several empirical studies have investigated the properties of the above validity coefficients, both in relation to one another and to traditional item selection statistics (Cox & Vargas, 1966; Crehan, 1974; Gorth & Hambleton, 1972; Henrysson & Wedman, 1973; Herbig, 1976; Kosecoff & Klein, 1974; Wedman, 1974b). On the whole, it appears that pretest-posttest item validation results in a more or less different item selection than does traditional statistics computed from the posttest data. Occasionally, however, these differences are of the same order as the mutual differences between the pretest-posttest coefficients. Most of these studies also try to detect clusters of coefficients and use correlation techniques for this purpose without realizing that most clusters can be expected a priori because of spurious correlations (the coefficients are defined on the same proportions from Table I). Crehan (1974) evaluates a number of pretest-posttest coefficients in view of their possible contributions to improving mastery decisions, and Smith (Note 1) does the same not using empirical but simulated data. The fact that he chooses the Rasch model to simulate the data and therewith represents items having equal discriminating power might account for his not finding any trend. Pretest-posttest coefficients are also used in other studies to assess the quality of item-writing techniques (Roid & Haladyna, 1978) and to evaluate mathematics instruction (Romberg & Braswell, 1973).

If one wishes to interpret the above coefficients against the background of some population from which the examinees are sampled or to take account of the measurements error inherent in the item responses, tests of statistical significance are needed. Brennan and Stolurow (1971) and Brennan (1972) discuss some tests that can be so used. From the equivalence of (1) and the numerator of (7), it follows that McNemar's chi squared can serve as a statistic to test the significance of D_{pp} . Nicolich, Nicolich, and Raph (1974) give a significance test for their average item change rating, which, as we pointed out earlier, is in fact identical to D_{pp} . A discussion of the Lazerfeld-Kendall turnover model in Wilcox (1977) is of interest for the problem of estimating the proportions in Table I and the above coefficients.

Some Criticism of Pretest-Posttest Validation

All validity coefficients we have considered so far are based on the same idea of instructional sensitivity, require pretest-posttest administration, and entail gain scoring. Although in many papers these features are welcomed as being typically criterion referenced, these coefficients have many disadvantages and serious interpretation problems. We shall show this below and observe that, to our knowledge, most of them have not hitherto been noticed.

Maximizing Posttest Scores

Our first criticism is the effect of item selection by the pretest-posttest method on the distribution of posttest scores. Suppose we use the D_{pp} coefficient and strive after items with a high D_{pp} value. The maximum D_{pp} value is reached for $p_1 = 0$ and $p_{.1} = 1$, that is, for a maximum p value for the posttest. Unlike traditional item analysis,

selection according to the D_{pp} coefficient has a one-sided effect on the posttest score distribution: All posttest scores are maximized and in the ideal situation the distribution is as shown in Figure 1a. Since in criterion-referenced or mastery testing the pass-fail decisions are made with a fixed percentage of items correct as cut-off scores and are not based on a fixed percentage of passes, as sometimes is the case in norm-referenced testing, the effect of D_{pp} item analysis on the percentage of passes is clear. The higher the D_{pp} values of the items in the final test, the higher the percentage of passes for any given cut-off score smaller than the maximum possible score. This seems to be an undesirable result. The final test loses its feedback to both students and instructors. The instruction confirms itself, as it were, via the item analysis, and the test is no longer the objective device students and instruction are tested with. Especially when pretest-posttest coefficients are used for assessing the quality of instruments for program evaluation (e.g., in Barta, Ahn, & Gastright, 1976), this methodological circular reasoning is seriously at fault. As an aside, note how traditional item analysis also has a predictable effect on pass-fail decisions. Selecting items with a p value around .50 and maximum item-test correlation, as is usual in norm-referenced measurements, leads to a score distribution as shown in Figure 1b. That this dichotomy arises when maximizing item-test correlation has sometimes been recognized, for instance, in the attenuation paradox discussion of the 1950s (Lord & Novick, 1968, p. 344). It has not been recognized, however, that for any cut-off score not equal to the maximum or minimum possible score it implies 50 percent of passes.

Validity of Instruction

A second weakness of the pretest-posttest item validation method has been noted earlier (e.g., Herbig, 1975, 1976; Wedman, 1973; Smith, Note 1). Pretest-posttest item

FIGURE 1. Ideal score distributions pertaining to three different item-selection methods: (a) pretest-posttest, (b) classical, and (c) Guttman analysis

validation only has meaning when the validity of the instruction is guaranteed; that is, when the actual learning effects coincide with the effects formulated in the learning objectives. When the validity of instruction is not established, low pretest-posttest may be due to poor instruction rather than weak test items. In fact, the pretest-posttest method of item validation requires a complete learning experiment in which the instruction and learning processes are fully controlled and tuned to the learning objectives with respect to which the items ought to show validity. Such requirements of control, however, are even too high for many learning experiments and unrealistic for the common practice of education. The hypothesis of instruction validity involves complex systems and processes and is much harder to verify than the hypothesis of the validity of a relatively simple item. Merely for this reason it is preferable to seek another method for validating criterion-referenced items.

Threats to Validity

Recognizing the pretest-posttest approach as an experimental method brings us to another criticism of this validation method. In the Campbell and Stanley (1966) notation the pretest-posttest method is a quasi-experiment $O_1 \times O_2$ design which, as is known, is liable to several possible factors jeopardizing its internal validity ("validity" now taken not in the psychometric but in the experimental methodological sense). Threats to validity that may make the results of the pretest-posttest method of item validation uninterpretable are history (the specific occurrences during the instruction that are not inherent to it but that may influence one or more of the items), maturation (e.g., learning that results from other instructional units parallel to the unit used for item validation), testing (the pretest administration that makes students alert to some parts of the instruction and by which the supposed sensitivity of items to instruction is in fact sensitivity of the instruction to the items), experimental mortality (dropout of students during the instruction, which may bias the D_{pp} values), and all possible interactions between these factors.

When groups of experts and nonexperts are used instead of preinstruction and postinstruction groups, the design is, in Campbell and Stanley's terminology, a Static-Group Comparison Design:

$$\frac{X \quad O_1}{\quad \quad O_2},$$

with X denoting a treatment and O an observation.

In addition to some of the confounded factors mentioned above there may be a selection effect involved in that, apart from the intended expertise denoted by X , both groups may differ in other important ways. When using a fallible indicator for selecting both groups, we may also expect a regression effect that biases the resulting D_{pp} value. All these factors are threats to validity in the double meaning of the word: They threaten the internal validity of the experiment, and in doing so they yield item validity coefficients that are, as such, not interpretable.

Course-specific Items

Cronbach (1963) has called attention to test items that are too course-specific. Generally, an item is course-specific if it uses terminology, symbols, notational

conventions, and the like taught in the course but irrelevant to the knowledge and skills the item is supposed to test. Selecting items according to their D_{pp} value may imply that we capitalize on course-specific learning and that, as a result, the remaining items still do not possess validity, even when the instruction would be completely valid and none of the preceding threats to validity is present. In a somewhat different context, Bereiter (1963) has also argued that selecting items on the basis of gain may produce unexpected results in that this method makes it possible to capitalize on specific intervening experiences.

The last few points of criticism show that, in general, D_{pp} values represent uninterpretable outcomes. Without further evidence it is unclear whether $D_{pp} = 0$ indicates an item that is completely insensitive to instruction (invalid) or that the instruction was ineffective or some of the threats to validity have biased the D_{pp} value downwards. Similarly, $D_{pp} = 1$ may indicate an item that is successful from the pretest-posttest method point of view but also a history, maturation, or testing effect or to capitalization on the course-specificity of the item formulation. Apart from these undesirable interpretation problems, more points of criticism can be raised.

Implicit Weighting

Although D_{pp} is defined in formula (1) with the help of the marginal proportions from Table I, it can also be written as a function of the joint proportions:

$$D_{pp} = p_{01} - p_{10}. \quad (8)$$

This form illustrates the close correspondence between D_{pp} (and its Brennan-Stolurow modification in [2]) and the pretest-posttest coefficients in (3), (5), (6), and (7). This form also illustrates that D_{pp} is in fact an index based on the two proportions of "turnover" (transition from one score category to the other). It is important to note that (8) uses an implicit weighting and that, with the exception of the sign, both proportions of turnover have the same weight: A transition from item score 1 to 0 can be compensated by a transition from 0 to 1. Setting aside the possibility of guessing (which we shall deal with below), for an item being constructed to record an increase in behavior and administered in a situation supposed to produce this increase, it seems strange to record a decrease. A much heavier penalty would be appropriate. Similarly, the equal weights of zero for the proportions p_{00} and p_{11} implied by (6) seems to be undesirable. The latter points to the proportion of students, who probably because of previous instruction are able to produce the correct response on the pretest and are still able to do so on the posttest. This is in accordance with the view underlying the pretest-posttest method. The former, however, points to students who produce the incorrect response on the pretest and, in contrast, still produce the incorrect response on the posttest. A more negative weight would also be more appropriate here. It is possible to use an explicit weighting in which the weights are specified according to the specific circumstances of the situation, for example, by defining a weighted sum of proportions scaled on the standard interval [0, 1].

Guessing

When true-false or multiple-choice items with only a few alternatives are used,

unexpected item score patterns can be explained by guessing. This suggests the application of a correction for guessing according to the knowledge or random guessing model:

$$p_{ij} = c + (1 - c)p'_{ij}, \quad (9)$$

with p'_{ij} denoting the corrected proportion and c the probability of success under random guessing. It is easy to see that this correction results in a chance corrected D_{pp} equal to $D_{ppc} = (p_{.1} - p_{1.})/(1 - c)$, and that this changes the order of the items only when these differ in their c value. As a first approximation, for q -choice items $c = q^{-1}$ can be used. It should be remembered, however, that applications of latent trait theory generally show estimates of c with a value somewhat lower than q^{-1} .

Population Dependency

Popham and Husek (1969) have made a plea to reject the classical test model for analyzing criterion-referenced measurements and to replace it by a model which is better suited to their peculiarities. According to them, classical test theory relies too much on variance to be usable for the mostly "range-restricted," criterion-referenced scores. For item analysis this implies that, for example, the point biserial item-test correlation is unusable, and, in this connection, they refer to Cox and Vargas' D_{pp} as the precursor of a new, criterion-referenced test theory. In discussing this argument, we have pointed out elsewhere that the classical test model is almost a tautology and that it (apart from the trivial requirement that the test score variance must be larger than zero to prevent some expressions from being indeterminate) does not contain any assumption saying that the score distribution displays much variability (van der Linden, 1977a). Thus, it is certainly not correct to conclude that a small amount of variability of test scores impairs the "fit" of the classical test model and that this is a reason for rejecting the model. However, for other reasons, Popham and Husek's analysis is not radical enough. It is known that test and item parameters of the correlation type are variance dependent and thereby show a range restriction effect. When interpreting values of these parameters, this should always be taken into account, whether there is large (!) or small variability in the test scores. It therefore seems desirable to have variance independent test and item parameters (not just parameters insensitive to distributions with a small variance), or, better, still, parameters that are independent of any distributional characteristic. Item p values satisfy this requirement no more than the classical correlation-based parameters, and the D_{pp} coefficient is, as the other pretest-posttest coefficients discussed above, population dependent as well. Brennan and Stolurow (1971) and Herbig (1975, 1976) must have felt this when formulating coefficients (2) and (6) to arrive at a pretest-posttest index independent of the distributions of pretest scores. For a coefficient intended to characterize the properties of an item, it is unacceptable that it is a function of the capacities of the examinees on the pretest. A pretest distribution also reflects the preceding history of the examinees and not only qualities of the item. The same, however, applies to the posttest. The effect of the instruction can on many points be less or more (!) than is formulated in the objectives. This shows itself in the posttest score distribution, and it is incorrect that our impression of the qualities of the item also depends on these properties of the posttest distribution. If, following Brennan and Stolurow or Herbig, we try to make coefficients (2) and (6) independent of the

posttest distribution, then we have to look for a coefficient registering an increase in item p value but independent of the pretest and posttest p value. This contradiction points to the incompatibility of the requirement of population invariance and the pretest-posttest method. Note, however, that population invariant parameters can be derived from latent trait theory.

Content Validity of Scores

Elsewhere we have indicated that for criterion-referenced measurements to be valid three hierarchical requirements must be met (van der Linden, 1979). The first requirement is that a mastery continuum formulated in the learning objectives must underlie the items and the test. The second requirement is that this continuum must have a known direction so that it is clear which scores point to a high and which to a low level of mastery. The third requirement—and for criterion-referenced measurement the most important one—is that for each relevant score or point of this continuum it must be known what its empirical reference is, with what domain of possible behavior it corresponds. The first two requirements together determine what is called the *validity of the test*. All three requirements together determine what could be called the *content validity of scores* and are requirements that can be derived from well-known definitions of criterion-referenced measurements (e.g., in Ebel, 1962, Glaser, 1963; and Glaser & Nitko, 1971). Note that both first requirements also apply to norm-referenced measurements; the last one, however, does not. In norm-referenced measurement the concern is with the relative position of examinees in a distribution of scores on a known continuum with a known direction but not with a behavioral interpretation of these scores. Although the pretest-posttest method has been introduced as a method for validating criterion-referenced items, it only meets the first two requirements. Given valid instruction and a fully controlled learning situation, an increase in item p value can with some reason be interpreted as an indication that the item indeed covers the intended mastery continuum. But this method does not meet the third and most critical requirement. The information that a test consists, for example, of 20 items with high D_{pp} values does not yield any clue to the empirical contents of a score of, say, 16 items right or the domain of behavior an examinee with this score is able to master. Paradoxically, a method introduced for validating criterion-referenced items seems to be suited only for norm-referenced purposes.

Local Discrimination

Criterion-referenced measurements are used for mastery decisions; that is, for decisions on whether a student exceeds some minimal level of mastery, θ_c (Hambleton & Novick, 1973; van der Linden, 1980). (There is also a state conception of mastery that conceptualizes mastery as a latent class underlying the items and not as a minimum level on a continuum. For a discussion, see Meskauskas, 1976.) D_{pp} is supposed to be a (criterion-referenced) discrimination index because it reflects the degree to which the item is sensitive to the transition of examinees from the “no mastery” to the “mastery area” on the mastery continuum. A criterion-referenced test is better suited for this purpose as its items have high discriminating power in the neighborhood of θ_c .

A high D_{pp} value is no guarantee that the item indeed possesses discriminating power at θ_c , however. Two items with the same difference between pretest and

posttest p value may cover different intervals of the mastery continuum with different degrees of discrimination. In order to be able to conclude whether an item has a high discriminating power at θ_c , it must be known how the probability of a correct item response varies with the level of mastery and where θ_c is located. No more than a suitable validation method of the pretest-posttest method can therefore be considered a suitable method for determining the discriminating power of criterion-referenced items.

A Latent Trait Look

Using one (external) criterion, the pretest-posttest method tries to accomplish two things at the same time, the assessment of the validity and the discriminating power of criterion-reference items, but accomplishes neither. It seems better to carry out separate analyses, each with a different and suitably chosen criterion. Before going into details, however, we look at the pretest-posttest method from a latent trait point of view to pinpoint precisely what causes its flaws.

Latent Trait Analysis

In the preceding paragraphs it was assumed that a single (mastery) continuum underlies the items. Although occasionally the opinion can be met that heterogeneous item domains are usable as well (Millman, 1974), homogeneous item domains are generally considered as a requisite for a criterion-referenced interpretation of test scores (see, e.g., Hambleton & Gorth, 1971; Macready & Merwin, 1973; Popham & Husek, 1967). We denote this underlying mastery continuum by θ and the cut-off score on it that is used for mastery decisions by the value θ_c . Latent trait theory is concerned with how the probability of a successful item response varies as a function of θ and item properties like difficulty, discriminating power, and the degree to which the item allows that the correct response is produced by mere guessing. A model often chosen because of its flexibility is the three-parameter logistic model

$$P_i(+|\theta) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i)]\}^{-1}, \quad (10)$$

where $P_i(+|\theta)$ denotes the probability of a correct response to item i given θ and a_i , b_i , and c_i are interpreted as the discriminating power, difficulty and guessing parameter of item i (Birnbau, 1968, pp. 399–405). Equation (10) with $c = 0$ is known as the two-parameter logistic model. It assumes that no guessing is possible but that the items may differ in difficulty and discriminating power. Another popular model, mainly because of the fact that the number right score and item p values are sufficient statistics for estimating its parameters, is the Rasch model. It results from (10) by putting $c_i = 0$ and $a_i = 1$

$$P_i(+|\theta) = \{1 + \exp[-(\theta - b_i)]\}^{-1} \quad (11)$$

(Rasch, 1960).

$P_i(+|\theta)$ can be conceived as the *local* p value for item i . For each value of θ , it gives the expected proportion of correct responses. The ordinary item p value for a population with probability function $g(\theta)$ is obtained from (10) or (11) by

$$p_i = \int_{-\infty}^{+\infty} P_i(+|\theta)g(\theta)d\theta. \quad (12)$$

Let $g_0(\theta)$ and $g_1(\theta)$ denote the pretest and posttest distribution of examinees on the mastery continuum. It then follows for Cox and Vargus' coefficient that

$$D_{pp} = \int_{-\infty}^{+\infty} P_i(+|\theta)[g_1(\theta) - g_0(\theta)]d\theta. \quad (13)$$

(The other pretest-posttest coefficients can be analyzed similarly, using the joint pretest-posttest distribution $g_{01}(\theta, \theta)$ in addition to the marginal distributions $g_0(\theta)$ and $g_1(\theta)$.) Note that D_{pp} , in the form given in (13), consists of two parts: the item characteristic curve and the difference between the posttest and the pretest mastery distribution. It is the latter that causes many of the problems mentioned in the preceding section. The difference between both distributions has nothing to do with any property of the item but is sensitive to all (valid and invalid) aspects of the instruction and the confounding factors mentioned earlier. It makes D_{pp} population dependent and raises interpretation problems. The former part, however, is independent of any distributional characteristic and reflects only properties of the item. Moreover, it can be estimated on the basis of a single test administration; this prevents arbitrary weighting problems as with D_{pp} . It therefore seems wise to use the item characteristic curve, particularly its properties in the neighborhood of the mastery score θ_c , for criterion-referenced item analysis.

Analysis of CR Items

The first requirement in analyzing criterion-referenced items is a mastery score or standard with a known and wanted behavioral interpretation. Although there are several procedures helpful in selecting mastery scores, as for instance "Counting backwards from 100%," the Nedelsky method, the kernel item method, and Huynh's (1970) referral task method (for an overview see Glass, 1978), all mastery score setting is ultimately based on an arbitrary judgment or decision. This point has been stressed by several authors (e.g., Hambleton, 1978; Jaeger, 1976; Linn, 1978; Popham, 1978; Shepard, 1976). In Huynh's method, success on a referral task is used as an external criterion for setting the mastery score, and the latter can therefore be interpreted using the former. In the kernel item method judges indicate which items represent the mastery level best. Next, the mastery score is computed from the statistics of these items (de Groot & van Naerssen, 1975, sect. 19.4). With the Nedelsky method, experts directly judge the item options with a view to the learning objectives, while the method of counting backwards supposes a domain of items, judged to be in correspondence with the learning objectives, for which a certain expected proportion of correct item responses is needed to be qualified as a master. These and all other methods of selecting mastery scores yield cut-off scores on the (classical test model) true-score underlying the item domain or test. However, to perform a latent trait analysis, a cut-off or mastery score on the latent continuum is needed. It is known that the relative true score, T , and the latent continuum, θ , are related by

$$T = \bar{P}(\theta) \equiv \frac{1}{n} \sum_{i=1}^n P_i(+|\theta), \quad (14)$$

where n is the number of items in the test or the domain (Lord & Novick, 1968, sect. 16.13.3). This relation is known as the test characteristic curve. (For domains of

items, it seems more appropriate to call this the domain characteristic curve.) Since the test characteristic curve is monotone increasing for the models discussed above, it determines a unique value of θ for each value of T , and can be used to transform the relative true mastery score, T_c , into the latent mastery score, θ_c . Note, however, that because of the summation involved, it is only possible to compute T from θ . The procedure is, therefore, to select a trial value for θ_c and to compute the associated value of T_c via (14). If the latter is smaller (larger) than the true mastery score, a larger (smaller) value for θ_c is selected until sufficient accuracy is attained. For a large number of items, this may entail the necessity of a (short) computer program. But fortunately most of the available latent trait analysis computer programs standardly output the test characteristic curve, and this can be used to simply read θ_c off. (See the example below.)

Once we have a mastery score, this can be used as an internal criterion for evaluating the discriminating power of the items. For criterion-referenced items, it is not sufficient to be discriminating "at large." They are used to separate masters from nonmasters, and the larger their discriminating power around θ_c the better they do this. With the item characteristic curve, this means that the items must show a high probability of a successful response for θ values just to the right of θ_c and a low probability for values just to the left of θ_c . In the ideal case the items have Guttman characteristic curves displaying their jump over θ_c :

$$P_i(+|\theta) = \begin{cases} 1 & \text{for } \theta \geq \theta_c \\ 0 & \text{for } \theta < \theta_c \end{cases} \quad (15)$$

Compare the score distribution resulting from a test of n of these Guttman items with the two distributions given in Figure 1. Denoting the number of masters ($\theta \geq \theta_c$) by m , this score distribution is depicted in Figure 1c. Now, again supposing a cut-off score not equal to the maximum or minimum value of the observed score variable, there is no a priori determined proportion of passes but a proportion that is equal to the actual proportion of masters among those who have taken the test. This is exactly what we want a mastery test to show. Guttman items have only theoretical meaning, since in practice we do not meet item characteristic curves possessing this form. The characteristic curve given in (15) is, however, important as an ideal in view of which we select criterion-referenced items for mastery decisions. In another paper, we have proposed to look at the first derivative of the item characteristic curve to θ in the point θ_c .

$$\frac{\partial}{\partial \theta} P_i(+|\theta)]_{\theta=\theta_c}, \quad (16)$$

the idea being that (16) represents the slope of the item characteristic curve for $\theta = \theta_c$ and that items with a steep slope at this point come close to the ideal formulated in (15) (van der Linden, 1977b). Guessing not being possible ($c_i = 0$), the item-discriminating power parameter a_i given in (10) can be viewed as a special case of (16), since a_i is then proportional to the value of the first derivative for $\theta = b_i$.

Note that (16) also has an interpretation that is analogous to the D_{pp} statistic. Using the definition of the first derivative of a function, (16) can be written as

$$\lim_{\theta_1, \theta_r \rightarrow \theta_c} \frac{P_i(+|\theta_r) - P_i(+|\theta_1)}{\theta_r - \theta_1}, \quad (17)$$

in which θ_1 and θ_r are values of θ to the left and to the right of θ_c , respectively. D_{pp} is the difference between the posttest and pretest item p value, while (17) is a function of the difference between the local p value of a master and a nonmaster arbitrarily close to θ_c . Van Naerssen (1977a) has at the same time proposed an index for selecting criterion-referenced items that is the first derivative of the expected *true score* given an observed score, $E(T|x)$, to x . In a second paper (van Naerssen, 1977b) he relates this to the slope of item characteristic curves and gives graphs for estimating the latter.

The index given in (16), however, has an unexpected property that makes it less appropriate for selecting criterion-referenced items. This appears when looking at the optimal test length. Analogous to (16), a test should be optimally discriminating in θ_c when the first derivative of the test characteristic function, $\bar{P}(\theta)$, to θ_c is maximal. $\bar{P}(\theta)$ is the average of the characteristic curves $P_i(+|\theta)$ of the items in the test, and the first derivative of $\bar{P}(\theta)$ is thus equal to the average of (16) taken over the test. But since an average cannot be larger than its largest term, it follows that an optimal discriminating test should always consist of only one item, namely the item with the largest value for (16). Clearly, this is a paradoxical result not only going against our intuition but also against what we know from psychometric research on test length. The reason is that, restricting ourselves to the slope of $P_i(+|\theta)$, we overlook an important factor influencing the local discriminating power: the scatter of the item responses at $\theta = \theta_c$. The larger this scatter, the less the discriminating power of the item in this point. Since the item responses constitute a dichotomy, the scatter at θ_c is equal to $P_i(+|\theta_c)[1 - P_i(+|\theta_c)]$. Combining this and (16) into one index naturally leads to an expression which is equal to the value of the Birnbaum item information function (Birnbaum, 1968, p. 454; Lord, 1980, sects. 5.4 and 11.11) at $\theta = \theta_c$.

$$I_i(\theta_c) = \frac{P_i'(+|\theta_c)^2}{P_i(+|\theta_c)[1 - P_i(+|\theta_c)]}, \quad (18)$$

where $P_i'(+|\theta_c)$ denotes the derivative given in (16). Formula (18) is the item information for the three-parameter logistic model; for the Rasch model this reduces to the form

$$I_i(\theta_c) = P_i(+|\theta_c)[1 - P_i(+|\theta_c)]. \quad (19)$$

Information functions are well anchored in statistical theory (see, e.g., Lindgren, 1976, sect. 4.5.4). One useful property is that in M.L. estimation the standard error of estimation is equal to $1/\sqrt{I}$.

For the purpose of this paper it is important to note that $I_i(\theta_c)$ has none of the earlier mentioned failings that D_{pp} is subject to. All these failings emanate from the fact that D_{pp} is a dual administration index based on population-dependent p values. $I_i(\theta_c)$, on the other hand, can be computed from a single test administration and uses only the population-invariant item characteristic curve.

It remains to indicate how item information functions can be used for item analysis purposes. Generally, item analysis ought to result in a test which is as most informative in θ_c as possible. For the two- and three-parameter model, the test information is the sum of the information of the item when the test is scored using locally best weights (Birnbaum, 1968, sect. 20.3). In that case we can simply gather the items with the largest value for (18) until we have reached a number predeter-

mined on practical grounds or the test information in θ_c exceeds some predetermined value. If no weighted scoring is used, the best strategy seems to compute the test information with and without the item, and to use the amount of information added by the item to the test as a selection criterion. For the Rasch model, weighted scoring is no condition for an additive relation between test and item information, and the items can be selected directly according to (19).

It should be realized that the estimated values of the item information function can only be compared with each other; they do not provide information on an absolute scale. For a single information function it is always possible to take any desired value by applying a suitably chosen transformation to the latent variable θ (Lord, 1975).

An Empirical Example

An empirical study was done to compare the D_{pp} coefficient and the item information function $I(\theta_c)$. The purpose of this study was to illustrate, for an educational situation typical of pretest-posttest validation, how item analysis according to D_{pp} and $I(\theta_c)$ values can lead to different selection of items.

The instructional unit in this study was the unit "Forces and Motion" from a course in physics in which 10th-grade pupils from general secondary schools were introduced into elementary mechanical concepts as mass, inertia, speed, acceleration, and two of the laws of motion. This relatively homogeneous unit, which took 16 consecutive classes of 50 minutes each, was prepared by the Project Team Curriculum Development Physics (PLON) of the State University at Utrecht, The Netherlands. The test items, all of the three- and four-choice type, were written by professional item writers of the National Institute of Educational Measurement (CITO), Arnhem, The Netherlands, and were selected by the PLON team. Altogether, 182 pupils from different schools but all receiving the same instruction and using the same learning material participated in this investigation. Some of them dropped out after the pretest, however. (For a discussion of sample sizes required for latent trait analysis, see below.) The pretest-posttest design was realized by administering the same 25-item test 4 weeks before the first and shortly after the last class. Some descriptive results of the pretest and posttest are given in Table II (SEM is the standard error of measurement and N the number of subjects). It is surprising to see how much the pupils knew about the course beforehand and how little the instruction added. The low variability in test scores at both administrations may account for the low KR-20 values; the standard error of measurement tells more about the measurement accuracy of the test.

Attempts to analyze the 25 items with the Birnbaum model given in equation (10) must be abandoned because the available computer program failed to show conver-

TABLE II
Descriptive Statistics Pretest and Posttest Administration

	Mean	Var	KR-20	SEM	n	N
Pretest	12.80	4.24	.05	2.01	25	182
Posttest	14.46	5.38	.32	1.92	25	156

gence on a large number of items. We therefore used the simpler Rasch model (equation 11) and the computer program BICAL (Wright & Stone, 1979) to analyze the posttest data. Seven items showed a bad fit to the Rasch model and were eliminated; a reanalysis of the remaining 18 items is given in Table III. The right-hand column gives the values for the BICAL item fit coefficients. All items have fit values below four times the standard error (.12) above their expected value of one, as is recommended in the BICAL manual, with the exception of items 4 and 9. However, these two showed a good fit with values $V_4 = .90$ and $V_9 = 1.36$ in the run with the complete collection of items. They were therefore retained. The overall fit of the 18 items in Table III was equal to 1.026, which is approximately one standard error (.028) above its expected value of one. These 18 items were used in our illustration of the differences between the D_{pp} coefficient and the item information function. As an aside, we note that for a real item analysis we certainly do not recommend to further neglect items not fitting the latent trait model. This would only be appropriate if a sufficient number of similar items were available. In general, however, the best strategy is to examine the items for possible flaws and to rewrite and reanalyze them until they fit both the model and the objectives intended to measure.

It appears that the pretest and posttest item p values in Table III have large variability but that the posttest-pretest difference, D_{pp} , is generally small and a few times even negative. D_{ppc} is D_{pp} corrected for guessing using the knowledge or random guessing model. Although the D_{ppc} values are, as could be expected, higher than the corresponding D_{pp} values, their rank order is virtually exactly the same (see Table V further on).

TABLE III
Several Item Statistics Computed for the 18 Rasch Items

Item	p_1	p_2	D_{pp}	D_{ppc}	b_i	$I_{.42}$	$I_{2.07}$	V_i
1	.69	.74	.05	.07	-.28	.22	.08	.53
2	.06	.10	.04	.06	2.81	.08	.22	.68
3	.82	.94	.12	.18	-1.83	.09	.02	.58
4	.32	.49	.17	.25	.75	.24	.17	2.08
5	.86	.95	.09	.13	-2.06	.07	.02	.28
6	.86	.85	-.01	-.01	-.91	.17	.05	1.15
7	.82	.88	.06	.09	-1.17	.14	.04	.70
8	.87	.93	.06	.09	-1.74	.09	.02	.99
9	.63	.71	.08	.12	-.16	.23	.09	2.49
10	.68	.87	.19	.28	-1.06	.15	.04	.84
11	.65	.75	.10	.13	-.34	.22	.08	.58
12	.13	.14	.01	.01	2.40	.11	.24	1.37
13	.68	.83	.15	.20	-.78	.18	.05	1.09
14	.08	.21	.13	.17	1.98	.14	.25	.83
15	.23	.24	.01	.01	1.77	.16	.24	1.46
16	.55	.62	.07	.09	.25	.25	.12	.89
17	.93	.94	.01	.01	-1.94	.08	.02	.53
18	.25	.15	-.10	-.13	2.31	.14	.25	1.41

Table III contains two series of item information function values, one for $\theta_c = .42$ and the other for $\theta_c = 2.07$. To obtain the first latent mastery score, a Nedelsky procedure was used. Nine members of the PLON-CITO team and a high school teacher were willing to participate; the results for the 18 Rasch items are displayed in Table IV. As noted earlier, the Nedelsky procedure yields an expected observed or true score that has to be transformed into a latent score via the test characteristic curve equation (14). The program BICAL outputs for each integer score this latent score value, however, and the latent mastery score corresponding with $T_c = 10.49$ was obtained by linear interpolation from $\theta = .24$ ($T = 10$) and $\theta = .60$ ($T = 11$). Also for the purpose of illustration, a second latent mastery score was computed. This was $\theta_c = 2.07$ obtained similarly from $T_c = 14.40$. The latter corresponds to a relative true score of 80 percent for an 18-item test, a well-known cut-off score in the mastery learning literature. The values for the item information function in Table III were obtained by substituting the relevant θ_c and the estimated item difficulty parameter value b_i in the Rasch model equation (11) and using the resulting $P_i(+|\theta)$ values for computing $I_i(\theta_c)$ from (19). These simple computations can be performed on a hand calculator with exponential function.

Table V gives the Spearman rank correlation between the pretest-posttest coefficients and the item information function for $\theta_c = .42$ and $\theta_c = 2.07$. These values are low, as might be expected. For the sake of completeness the mutual correlations between D_{pp} and D_{ppc} and both information functions have been added. The former is virtually 1.00, indicating that in this particular case a correction for guessing hardly changes the item ranks. The latter is necessarily low: with the Rasch model, different latent scores generally yield different information values.

Table VI is perhaps more informative in showing the differences between both item analysis methods. For a number of situations, it gives the overlap in item detection between the pretest-posttest and the latent trait method. So if we want our item analysis to detect, for example, the six worst items from the 18 items available, there would only be three common items for $\theta_c = .42$ and no more than one (!) for $\theta_c = 2.07$.

TABLE IV
Results Nedelsky Mastery Score Determination (n = 18)

	No. of Subjects	Mean	SD	Latent Ned. Score
Nedelsky Score	9	10.49	2.00	.42

TABLE V
Spearman Rank Correlation Between Item Statistics (n = 18)

	D_{pp}	D_{ppc}	$I_{.42}$	$I_{2.07}$
D_{pp}	—	—	—	—
D_{ppc}	.996	—	—	—
$I_{.42}$	.23	—	—	—
$I_{2.07}$	-.19	-.24	.37	—

TABLE VI
Overlap in Item Detection Between Both Approaches (n = 18)

No. of Eliminated Items	I_{A2}/D_{pp}	I_{A2}/D_{ppe}	$I_{2.07}/D_{pp}$	$I_{2.07}/D_{ppe}$
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	1	1	1	1
5	1	1	1	1
6	3	3	1	1
7	3	3	2	2
8	3	3	2	2
9	6	6	4	4
10	6	6	5	5

Concluding Remarks

The result in Table VI demonstrates the loss possible when the analysis of mastery test items is based on the pretest-posttest method. Items that discriminate poorly at the mastery score θ_c still have a large probability of being selected by the pretest-posttest method. We have already indicated several weak points of the pretest-posttest method and all these may account for the lack of overlap in item selection. Most of them are summarized in the analysis given in equation (13): The pretest-posttest method mixes up two sources of information—the characteristics of the item and the differences between the pretest and posttest mastery distributions—and blames the former for the peculiarities of the latter. By doing so, it is not surprising to find that it weeds out items of high quality.

As noted earlier, the item information function does not mix up these two sources of information. Therefore, from a theoretical point of view, it is by far superior to the D_{pp} index. Looking at the practical aspects of the two item analyses, it can be noted that the pretest-posttest method requires two test administrations, whereas the latent trait analysis requires only one. On the other hand, the latent trait analysis can only be done using a computer, whereas the D_{pp} index can mostly be calculated by hand. It should also be noted that the latent trait analysis can entail the necessity of larger sample sizes. The Rasch analysis in this paper was based on 156 subjects (Table II). For the greater part of the applications this number is sufficient. Many accurate applications with smaller sample sizes are known; Wright and Stone (1979) use a successful example throughout their book which is even based on a sample size of 35. For the Birnbaum model the situation is somewhat different. In our example with 156 subjects the computer program failed to yield convergent estimates. Most successful applications of the Birnbaum model reported in the literature use sample sizes several times larger than we had at our disposal (see, e.g., Lord, 1980). This seems to suggest that, if sample size is the critical factor in obtaining converging estimates, the item information function can only be used in combination with the Birnbaum model in large-scale applications or when systematic item banking occurs. We note, however, that when the population distribution of latent scores may be

assumed to be normal, a simple estimation method for the Birnbaum model can be used which requires samples of the same size as in classical item analyses (Urry, 1974). In our example, we were not able to use this method because the test data did not lend sufficient support to the assumption of normality.

Finally, though we hardly know situations in which criterion-referenced testing is not used for decision making, it may be worthwhile to point out that item information functions are also valuable as item analysis tools in situations where the interest is purely in the use of criterion-referenced tests for measurement purposes. We then aim at items not showing maximum information at some cut-off score but in the region in which the scores are expected. Techniques to do this have already been suggested (Birnbaum, 1968, chap. 20; Lord, 1980; Wright & Stone, 1979, chap. 6). The virtue of this application of item information functions also lies in the fact that it provides population-invariant test design based solely on the characteristics of the items.

Reference Notes

1. Smith, D. U. *The effect of various item selection methods on the classification consistency of criterion-referenced instruments*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.
2. Coulson, D. B., & Hambleton, R. K. *On the validation of criterion-referenced tests to measure individual mastery*. Paper presented at the annual meeting of the American Psychological Association, New Orleans, 1971.

References

- Barta, M. B., Ahn, U. R., & Gastright, J. F. Some problems in interpreting criterion-referenced results in a program evaluation. *Studies in Educational Evaluation*, 1976, 2, 193-202.
- Bereiter, C. Some persisting dilemmas in measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.
- Berk, A. (Ed.). *Criterion-referenced measurements: The state of the art*. Baltimore: The Johns Hopkins University Press, 1980.
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
- Brennan, R. L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 1972, 32, 289-303.
- Brennan, R. L. *The evaluation of mastery test items* (Final Report). Stony Brook, N.Y.: State University of New York, Department of Education, January 1974. (EDRS No. ED 092 593)
- Brennan, R. L., & Stolorow, L. M. An empirical decision process for formative evaluation. *Research memorandum no. 4*. Harvard CAI Laboratory, Cambridge, Mass., 1971. (EDRS No. ED 048 343)
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago, Ill.: Rand McNally, 1966.
- Carver, R. P. Special problems in measuring chance with psychometric devices. In B. Baxter (Ed.), *Evaluative research: Strategies and methods*. Pittsburgh: American Institutes for Research, 1970.
- Carver, R. P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 1974, 29, 512-518.
- Cox, R. C., Evaluative aspects of criterion-referenced measures. In W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction*. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

- Cox, R. C., & Vargas, J. S. *A comparison of item selection techniques for norm-referenced and criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February 1966. (EDRS No. ED 010 517)
- Crehan, K. D. Item analysis for teacher-made mastery tests. *Journal of Educational Measurement*, 1974, 11, 255–262.
- Cronbach, L. J. Evaluation for course improvement. *Teachers College Record*, 1963, 64, 672–683.
- Cureton, E. E. The upper and lower twenty-seven per cent rule. *Psychometrika*, 1957, 22, 293–296.
- de Groot, A. D., & van Naerssen, R. F. *Studietoetsen* (Vol. 2). 's-Gravenhage, The Netherlands: Mouton, 1975.
- Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, 22, 15–25.
- Ebel, R. L. Criterion-referenced measurements: Limitations. *School Review*, 1971, 79, 282–288.
- Edmonston, L. P., & Randall, R. S. *A model for estimating the reliability and validity of criterion-referenced measures*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972. (EDRS No. ED 065 591)
- Findley, W. G. A rationale for the evaluation of item discrimination indices. *Educational and Psychological Measurement*, 1956, 16, 175–180.
- Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, 18, 519–521.
- Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1971.
- Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, 15, 237–261.
- Gorth, W. P., & Hambleton, R. K. Measurement considerations for criterion-referenced testing and special education. *Journal of Special Education*, 1972, 6, 303–314.
- Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 1978, 15, 277–290.
- Hambleton, R. K., & Gorth, W. P. *Criterion-referenced testing: Issues and applications* (Report No. TR-13). Amherst: University of Massachusetts, School of Education, 1971. (EDRS No. ED 060 025)
- 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.
- Hambleton, R. K. et al. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, 48, 1–47.
- Harris, C. W., & Wilcox, R. R. Brennan's B is Peirce's Theta. *Educational and Psychological Measurement*, 1980, 40, 307–311.
- Harris, N. D. C. An index of effectiveness for criterion-referenced items used in pre-tests and post-tests. *Programmed Learning and Educational Technology*, 1974, 11, 125–132.
- Henrysson, S., & Wedman, I. Some problems in construction and evaluation of criterion-referenced tests. *Scandinavian Journal of Educational Research*, 1973, 18, 1–12.
- Herbig, M. Zur Vortest-Nachtest-Validierung Lehrzielorientierter Tests. *Zeitschrift für Erziehungswissenschaftliche Forschung*, 1975, 9, 112–126.
- Herbig, M. Item analysis by use in pretests and posttests: A comparison of different coefficients. *Programmed Learning and Educational Technology*, 1976, 13, 49–54.
- Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, 14, 5–10.
- Hively, W. et al. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project* (CSE Monograph Series in Evaluation No. 1). Los Angeles: University of California Center for the Study of Evaluation, 1973.
- Hively, W., Patterson, H. L., & Page, S. H. A universe-defined system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, 5, 275–290.
- 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)
- Ivens, S. H. *A pragmatic approach to criterion-referenced measures*. Paper presented at a joint session of the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 1972. (EDRS No. ED 064 406)
- Jaeger, R. M. Measurement consequences of selected standard-setting models. *Florida Journal of Educational Research*, 1976, 18, 22-27.
- Kosecoff, J. B., & Klein, S. P. *Instructional sensitivity statistics appropriate for objective-based test items* (CSE Report No. 91). Los Angeles: University of California, Center for the Study of Evaluation, 1974.
- Lindgren, B. W. *Statistical theory* (3rd ed.). New York: MacMillan, 1976.
- Linn, R. L. Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement*, 1978, 15, 301-308.
- Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, 9, 13-26.
- Lord, F. M. The "ability" scale in item characteristic curve theory. *Psychometrika*, 1975, 40, 205-217.
- Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1980.
- Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
- Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain-referenced testing. *Educational and Psychological Measurement*, 1973, 33, 351-360.
- 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, Wis.: University of Wisconsin, Research and Development Center for Cognitive Learning, 1975.
- McGuigan, F. J., & Peters, R. J. Assessing the effectiveness of programmed texts: Methodology and some findings. *Journal of Programmed Instruction*, 1965, 3, 23.
- Mellenbergh, G. J., & van der Linden, W. J. The internal and external optimality of decisions based on tests. *Applied Psychological Measurement*, 1979, 3, 257-273.
- Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, 46, 133-158.
- Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education*. Berkeley, Calif.: McCutchan, 1974.
- Nicolich, M., Nicolich, L., & Raph, J. *Demonstration of techniques for optimizing the use of criterion-referenced achievement tests in children enrolled in a day care center*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974. (EDRS No. ED 090 313)
- Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, 28, 95-104.
- Ozenne, D. G. *Toward an evaluative methodology for criterion-referenced measures: Test sensitivity* (Report No. CSE-R-72). Los Angeles, Cal.: University of California, Center for the Study of Evaluation, 1971. (EDRS No. ED 061 263)
- Popham, W. J. (Ed.). Indices of adequacy for criterion-referenced test items. *Criterion-referenced measurement: An introduction*. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.
- Popham, W. J. As always, provocative. *Journal of Educational Measurement*, 1978, 15, 297-300.
- Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, 6, 1-9.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmark Paedagogiske Institut, 1960.

- Roid, G. H., & Haladyna, T. H. A comparison of objective-based and modified-Bormuth item writing techniques. *Educational and Psychological Measurement*, 1978, 38, 19-28.
- Romberg, T. A., & Braswell, J. Achievement monitoring via item sampling: A practical data-gathering procedure for formative evaluation. *Journal for Research in Mathematics Education*, 1973, 3, 262-270.
- Roudabush, G. E. *Item selection for criterion-referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973. (EDRS No. ED 074 147)
- Rovinelli, R., & Hambleton, R. K. *Some procedures for the validation of criterion-referenced test items*. Albany, N. Y.: Bureau of School and Cultural Research, New York State Education Department, 1973.
- Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch*, 1977, 2, 49-60.
- Saupe, J. L. Selecting items to measure change. *Journal of Educational Measurement*, 1966, 3, 223-228.
- Shepard, L. A. Setting standards and living with them. *Florida Journal of Educational Research*, 1976, 18, 28-32.
- 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 investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, 15, 111-116.
- Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, 11, 263-267.
- Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, 34, 253-270.
- van der Linden, W. J. Criteriumgeoriënteerde toetsen: Afscheid van het klassieke testmodel? In *Congresboek Onderwijsresearchdagen 1976*. Amsterdam, The Netherlands: Stichting Onderwijsresearch, 1977. (a)
- van der Linden, W. J. Pretest-posttest validatie van criteriumgeoriënteerde toetsen. In *Congresboek Onderwijsresearchdagen 1977*, Amsterdam, The Netherlands: Stichting Onderwijsresearchdagen, 1977. (b)
- van der Linden, W. J. Criteriumgeoriënteerd toetsen. In E. Warries (Ed.), *Beheersingsleren een leerstrategie*. Groningen, The Netherlands: Wolters-Noordhoff, 1979.
- van der Linden, W. J. Decision models for use with criterion preferred tests. *Applied Psychological Measurement*, 1980, 4, 469-492.
- 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.
- van Naerssen, R. F. Lokale betrouwbaarheid: Begrip en operationalisatie. *Tijdschrift voor Onderwijsresearch*, 1977, 2, 111-119. (a)
- van Naerssen, R. F. Grafieken voor de schatting van de helling van itemkarakteristieken. *Tijdschrift voor Onderwijsresearch*, 1977, 2, 193-201. (b)
- Wedman, I. *Theoretical problems in construction of criterion-referenced tests* (Educational Reports Umea No. 3). Umea, Sweden: Umea University and Umea School of Education, 1973.
- Wedman, I. On the evaluation of criterion-referenced tests. In H. F. Crombag & D. N. M. de Gruijter (Eds.), *Contemporary issues in educational testing*. The Hague, The Netherlands: Mouton, 1974. (a)
- Wedman, I. *Reliability, validity, and discrimination measures for criterion-referenced tests* (Educational Reports Umea No. 4). Umea, Sweden: Umea University and Umea School of Education, 1974. (b)
- Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pastorok, & R. R. Wilcox (Eds.), *Achievement test items: Methods of study* (CSE Monograph No. 6). Los Angeles, Calif.: University of California, Center for the Study of Evaluation, 1977.

WIM J. VAN DER LINDEN

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

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

AUTHOR

WIM J. VAN DER LINDEN, Afdeling Toegepaste Onderwijskunde, Technische Hogeschool Twente, Postbus 217, 7500 AE Enschede, The Netherlands. *Specialization*: Educational measurement and research methodology.