On Problems Encountered Using Decision Theory to Set Cutoff Scores

Dato N.M. de Gruijter
University of Leyden

Ronald K. Hambleton
University of Massachusetts, Amherst

In the decision-theoretic approach to determining a cutoff score, the cutoff score chosen is that which maximizes expected utility of pass/fail decisions. This approach is not without its problems. In this paper several of these problems are considered: inaccurate parameter estimates, choice of test model and consequences, choice of subpopulations, optimal cutoff scores on various occasions, and cutoff scores as targets. It is suggested that these problems will need to be overcome and/or understood more thoroughly before the full potential of the decision-theoretic approach can be realized in practice.

The decision-theoretic paradigm for setting cutoff scores with criterion-referenced tests works in the following manner. In making mastery decisions on the basis of criterion-referenced test scores, two kinds of errors can result: passing examinees who are true nonmasters and failing examinees who are true masters (Hambleton, Swaminathan, Algina, & Coulson, 1978). A loss function is attached to the two types of errors, and then a cutoff score is chosen to maximize expected utility or to minimize expected loss with the available examinee test data. Among the researchers who have described the use of decision theory for setting cutoff scores are Hambleton and Novick (1973), Huynh (1976), Van der Linden and Mellenbergh (1977), and Van der Linden (1980).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 1, Winter 1984, pp. 1–8

0146-6216/84/010001-08$1.65

Despite the conceptual appeal of the approach to some researchers, the approach appears to have a number of problems at the implementation stage which have not been thoroughly discussed in the psychometric literature, with the significant exception of a recent paper by Glass (1978). In fact, an inadequate cutoff score can arise through the use of decision theory because of these problems. To date, these problems have not been encountered or discussed to any extent, perhaps because of the few actual applications of decision theory for setting cutoff scores. Still, in view of the apparent attractiveness of the approach and the number of advocates for the approach in the criterion-referenced testing literature, it is not unreasonable to expect an increase in the number of applications in the future. The purpose of this paper is to discuss five problems that arise when decision theory is used to set cutoff scores. The basic paradigm for setting optimal cutoff scores will be reviewed in the first section of this paper. In the subsequent sections, five problems with the decision-theoretic paradigm are discussed.

Despite the problems with decision-theoretic procedures that will be considered in this paper, it may be that these procedures are the most appropriate to use in some situations. It should be emphasized that it is not the authors' intention to compare decision-theoretic procedures for setting standards with other available methods nor to suggest that the five problems are unique to decision-

theoretic procedures. Several of the problems are encountered with other standard-setting procedures as well.

The Basic Paradigm for Obtaining Optimal Cutoff Scores[1]

For the moment assume that a standard has been set on the domain score scale, where domain scores are defined as the relative true scores (or true proportion-correct scores) in the content domain of interest. The content domain of interest may be narrow or broad, but it must be clearly defined to permit valid criterion-referenced measurements (Hambleton et al., 1978). Examinees with domain scores, π, equal to or larger than the standard, π_0, are considered to have mastered the relevant content; they thus should pass a test which measures the content domain. Examinees with domain scores $\pi < \pi_0$ are nonmasters and therefore should fail. In other words, the utility of passing a master exceeds the utility of failing a master: $U_p(\pi) > U_f(\pi)$ for $\pi \geq \pi_0$. In the same way, the utility of failing a nonmaster exceeds the utility of passing a nonmaster. Therefore, $U_f(\pi) > U_p(\pi)$ for $\pi < \pi_0$. With more realistic utility functions, $U_p(\pi)$ and $U_f(\pi)$ are determined first and next the standard π_0 may be obtained as the value of π for which $U_p(\pi) = U_f(\pi)$. Mellenbergh and Van der Linden (1981) have a divergent point of view. According to their Formula 2, the point of intersection may deviate from π_0, defined as the minimum level of acceptable performance. This discrepancy disappears, however, when π_0 is defined as the lowest value of π on the basis of which an examinee can be passed. In other words, when there are no measurement errors, π_0 gives the border between pass and fail decisions.

Test scores, unfortunately, give fallible information about π. One reason is that only a sample of test items from the content domain of interest can be administered to examinees. Also, there are measurement errors due to faulty test items and to examinee guessing behavior. Therefore, decisions must be made under uncertainty, and some incorrect classificatory decisions will result. In the decision-theoretic approach to the cutoff score problem the cutoff score that maximizes the expected utility is chosen.

Consider now the population of examinees for whom the test is intended. Assume that the density of π in the population is $g(\pi)$ and that the joint distribution of observed scores and domain scores equals

$$f(x,\pi) = f(x|\pi)\ g(\pi). \tag{1}$$

For a given cutoff score C_x on an n-item test the expected utility can be written as

$$U = \sum_{x=0}^{C_x-1} \int U_f(\pi) f(x,\pi) d\pi + \sum_{C_x}^{n} \int U_p(\pi) f(x,\pi) d\pi. \tag{2}$$

Consider the example below with simple utilities:

a = utility of passing a true master,
b = utility of passing a true nonmaster,
c = utility of failing a true nonmaster, and
d = utility of failing a true master.

Clearly, $a > d$ and $c > b$.

Let A be the proportion of examinees correctly passed, B the proportion of examinees incorrectly passed, C the proportion of examinees correctly failed, and D the proportion of examinees incorrectly failed. Then, for this particular example:

$$\begin{aligned} U &= aA + bB + cC + dD \\ &= (d - a)D + (b - c)B + (A + D)a + (C + B)c. \end{aligned} \tag{3}$$

Since the proportion of true masters, $A + D$, and the proportion of true nonmasters, $B + C$, are constant regardless of the chosen cutoff score, the problem becomes one of finding C_x that maximizes U', where

$$U' = (d - a)D + (b - c)B, \tag{4}$$

or finding C_x that minimizes

[1]For the purpose of this paper it is convenient to make a distinction between "standards" and "cutoff scores." Standards are set on domain score scales to identify "masters" and "nonmasters." Cutoff scores are set on test score scales in order to make pass-fail decisions.

$$R = l_{10}D + l_{01}B, \quad (5)$$

where $l_{10} = a - d$ is the loss associated with failing a true master and $l_{01} = c - b$ is the loss associated with passing a true nonmaster.

An analysis similar to the one above can be found in, for example, Mellenbergh, Koppelaar, and Van der Linden (1977). Any of a multitude of loss functions or utility functions can be chosen. A threshold loss function was used by Hambleton and Novick (1973). A linear loss function was suggested by Van der Linden and Mellenbergh (1977). Other functions were suggested by Huynh (1976) and Novick and Lindley (1978).

Two approaches have been suggested for defining $g(\pi)$. In the frequentist view, $g(\pi)$ is a population distribution. However, $g(\pi)$ also can be taken to represent prior knowledge with respect to an examinee (Hambleton & Novick, 1973). In this paper it is assumed that $g(\pi)$ is a population distribution. In other words, prior knowledge with respect to an examinee from the population will be equated with the population density.

Inaccurate Parameter Estimates

The determination of the standard π_0 on the domain score scale, and the loss ratio (or, more generally, the utility or loss structure), is no easy task. The estimation of $g(\pi)$ can also be a major problem. The problems associated with each will be discussed below.

In a special issue on standard setting in the *Journal of Educational Measurement*, Glass (1978) criticized the use of decision theory for setting cutoff scores because the approach (as well as several other approaches) depends on an arbitrarily chosen standard, π_0; it might be added that in case π_0 is defined as the value of π for which $U_p(\pi) = U_f(\pi)$, π_0 is arbitrary to the extent that there is uncertainty regarding the specification of $U_p(\pi)$ and $U_f(\pi)$ near the point of intersection. In response to Glass, others (e.g., Popham, 1978) agreed that standard setting is judgmental, but they have argued that arbitrariness in the sense of well-considered judgment is unavoidable. Moreover, in view of the desirability of using test scores to influence decision-making, cutoff scores are seen as inevitable.

However, if standards are arbitrary, why bother at all about optimal cutoff scores? Why not set C_x equal to $n\pi_0$ (or the first integer exceeding $n\pi_0$), where n is the number of test items? Two reasons against such a proposal can be offered. First, it can be argued that a chosen cutoff score should be as precise as possible, also under uncertainty. Second, in some situations the optimal cutoff score may differ considerably from the simple alternative. This is most likely to occur when the standard differs from the domain score mean in the population, when test reliability is low, and when the loss ratio is extreme.

Uncertainty with respect to the value of π_0, about which Glass (1978) was concerned, has been discussed within the framework of decision theory (De Gruijter, 1980). First, π_0 has been treated as another variable with an associated prior distribution. This distribution reflects the uncertainty of the standard setter with respect to π_0. Second, robustness studies have been proposed (De Gruijter 1980; Vijn & Molenaar, 1981). In a robustness study, regions are obtained where values of π_0 give the same cutoff score. Clearly, such a study results in additional relevant information, such as the consequences of a chosen cutoff score (e.g., how many examinees will fail?). Finally, it has become clear that there should be an attempt to improve the initial standard using additional information such as examinee results (De Gruijter, 1980; Shepard, 1979).

A major obstacle to the use of decision theory is the problem of specifying the utility or loss structure. Many researchers, including Glass (1978) and Hambleton et al. (1978), have commented on this problem. Huynh (1976) has suggested that the specification of the losses should include a numerical assessment of emotional and psychological consequences due to incorrect decisions and an assessment of effects of overlearning. In many situations, however, time losses would constitute the determining factor (Huynh, 1976). So, the loss in failing a master would depend on the time such a student spends in repetition and remedial work. Novick and Lindley (1978) have presented a general approach for determining utilities. Further, an

interactive program, CADA (Isaacs & Novick, 1978; Novick, Isaacs, & Dekeyrel, 1977) has been developed that assists in a consistent specification of utilities. With respect to losses, a robustness study can be performed in order to find out to what extent uncertainty might be harmful with respect to the value of the losses. De Gruijter (1980) and Vijn and Molenaar (1981) have proposed robustness studies where the loss ratio and the standard π_0 are jointly varied. However, research on the choice and robustness of loss functions is only beginning.

Many researchers have assumed that the population distribution $g(\pi)$ can be approximated by a (two-parameter) beta distribution, which has convenient properties from a mathematical point of view (Huynh, 1976; Mellenbergh, et al., 1977; Wilcox, 1977). The parameters of the beta distribution are usually estimated with data from a sample of examinees; but the beta distribution does not seem adequate in all applications. For example, when there is substantial guessing, a three-parameter beta distribution may be more suitable (Lord, 1965). Also, when the domain score distribution is multi-modal, the beta distribution is a poor choice because it cannot provide a close approximation. Therefore, it seems most desirable and necessary to have techniques to estimate $g(\pi)$ that do not make any assumptions with respect to the distributional form. Maritz (1966) suggested such a technique, which can be applied if it is assumed that $g(\pi)$ is discrete. Another approach to the estimation of $g(\pi)$ was suggested by Lord (1969). His Method 20 is based upon finding a continuous, smooth solution, $\hat{g}(\pi)$. An application of this method within the context of criterion-referenced measurement is given by Livingston and Wingersky (1979). The sample size needed for a stable solution, however, seems to be prohibitively large. It must be added that neither the Maritz (1966) nor Lord (1969) methods are being used with the decision-theoretic approach to setting cutoff scores at this time.

Choice of Test Model and Consequences

In order to apply decision theory, the form of $f(x|\pi)$ must be specified; in fact, without knowledge of this conditional distribution, $g(\pi)$ cannot be estimated (see the previous section). Generally, the binomial distribution is chosen (e.g., Huynh, 1976; Mellenbergh, et al., 1977). This distribution seems adequate when items in the content domain of interest have the same or nearly the same difficulty level. The binomial model also seems adequate when every examinee answers a different random sample of items from the item domain. In practice, neither assumption is met: It is common for items to range in difficulty within a domain (see, e.g., the Individualized Criterion-Referenced Tests Technical Manual, Educational Development Corporation, 1980), and seldom are examinees administered separate samples of test items. Individualized tests can be costly and/or difficult to administer, with the exception of computer-administered tests.

When the binomial model is inadequate, an approximation of the generalized binomial model can be used instead (Wilcox, 1977). A specific approximation has been proposed by Lord (e.g., Lord, 1965). This distribution is a function of the relative true score on the particular sample of chosen items in the test. So the frequency distribution may be written as $f(x|\zeta)$, where ζ is the relative true score. Unfortunately, ζ generally differs from the relative true score in the item domain, π. For example, the item sample constituting the test can be relatively easy, and in that case, $\zeta > \pi$. When the relationship between π and ζ is known, there is not a problem. The utilities defined on the domain score scale can be transformed to the relative true score scale of the test, or $f(x|\zeta)$ can be written in terms of π. The relationship between ζ and π, however, is seldom known.

Generally, the problem of differences between true score scales for different tests drawn from the same content domain has been neglected. There are exceptions. Brennan and Kane (1977) analyzed the problem of differences between test forms in terms of generalizability theory. Another possibility is to correct for differences between test forms, i.e., to equate the true-score scales of different tests. With respect to equating, item response theory (Lord, 1980) offers promising techniques. But how well they might work with small examinee samples is unknown at this time.

Choice of Subpopulations

In the determination of the optimal cutoff score with decision theory, the population distribution $g(\pi)$ plays a major role. Generally, the cutoff score will be lower for populations with a higher level of achievement. A problem arises when a population of examinees is divided into subpopulations. Alternatively, the problem may arise in the opposite direction. How inclusive should a population be? Should it include, for example, handicapped students, students who have already failed the test previously, accelerated students, and so on? The definition of the population or subpopulations will impact directly on the resulting cutoff score set with the aid of decision theory. For example, examinees might be classified into subpopulations according to sex. In general, when the distributions of π differ in the subpopulations, different optimal cutoff scores for these subpopulations will arise.

A further complication arises when different loss structures are used for different subpopulations, as in culture-fair selection (Gross & Su, 1975; Petersen & Novick, 1976). Perhaps the idea of different loss structures seems strange in connection with criterion-referenced measurement. However, the idea is implicitly present in Huynh (1976), who has incorporated success on a referral task in his loss function. It is quite possible that success on a referral task is not a function of the level of π alone. Novick (1980) has also argued that utility may be multiattributed. So, it is possible that utility is a function of π and another variable η, $U(\pi,\eta)$. In this case,

$$U(\pi) = \int U(\pi,\eta)dP(\eta|\pi) \tag{6}$$

could be defined, where $P(\eta|\pi)$ is the conditional distribution of η given π. Clearly, $U(\pi)$ may vary from one subpopulation to the next, as the joint distribution of π and η varies. In order not to complicate the discussion, it will be assumed that differences in $U(\pi)$ between subpopulations are negligible.

It generally seem undesirable to divide a population of interest into subpopulations for the purpose of setting a unique cutoff score in each subpopulation. There are two reasons for support of this position, however.

First, many variables might be used to divide a population into subpopulations (e.g., age, region, religion, race, socioeconomic status). The problem is that when a cutoff score is set for each subpopulation, examinee mastery status will depend upon the choice of variable or variables for sorting examinees into subpopulations. The larger the difference in achievement levels between subpopulations, in general, the greater the difference will be in the optimal cutoff score for each subpopulation. The greater the difference in cutoff scores, the more important subpopulation membership will be in determining mastery status. For important examinations, political pressure for the selection of one variable over another will be substantial, and there may be substantial difficulties in determining examinee subpopulation membership for all but the simplest variables such as age and gender. If multiple classificatory variables are used, one additional problem should be added: The sample size in each subpopulation may be too small to result in the stable determination of optimal cutoff scores.

Second, recall that the distribution of π in a subpopulation is often chosen as the prior distribution for all examinees from that subpopulation. An examinee, however, might successfully argue that he/she is not a typical member of that subpopulation. A similar argument has been used within the context of culture-fair selection by Novick and Ellis (1977). They supported the concept of differential treatments but argued that group membership should be rejected as a criterion for differential treatment in favor of an estimate of individual disadvantage.

Optimal Cutoff Scores on Various Occasions

In some situations (e.g. university courses) different randomly equivalent forms of a criterion-referenced test are available to assess mastery in a relevant content domain. These forms may be administered to different groups of examinees on different occasions. When a common cutoff score across tests is set (a not unreasonable approach to take), it is not possible to optimize decision making after each set of test results is known, for that

would, in general, result in a different cutoff score for each test. This point will be developed further in the next section. This section will discuss the situation in which a cutoff score is determined on the basis of the test results.

Even when the cutoff score is determined after the test administration, it is possible to obtain the same optimal cutoff score on various occasions, at least when test forms do not vary in difficulty level and when essentially the same population is being dealt with on each occasion. On the other hand, problems can still arise. Suppose examinees learn what the cutoff score is. It is possible that when the optimal cutoff score has been low, possibly unexpectedly, the next group of examinees learning this information will spend less time in test preparation. Perhaps the good students will work only hard enough to pass the test, based upon their knowledge of the earlier cutoff score. This will mean that $g(\pi)$ will shift downward in the new group and the old cutoff score may no longer be optimal. Now, a new value for the cutoff score that is optimal in the new situation might be obtained and used. Even if this is done, however, a loss may be experienced.

This point will be demonstrated under the assumption that the loss structure remains the same. For this demonstration the loss structure introduced earlier as an example will be used. Assume that Equation 3 reflects the situation on the first test occasion and that

$$U^* = (d - a)D^* + (b - c)B^* + (A^* + D^*)a + (C^* + B^*)c \quad (7)$$

is the situation on the second test occasion. The contribution of

$$(A^* + D^*)a + (C^* + B^*)c \quad (8)$$

does not depend on the cutoff score chosen. It may, therefore, be neglected in the search for the optimal cutoff score on the second occasion. It is, however, important for a comparison between occasions. The proportion of true masters has decreased on the second occasion because the examinees have not worked as hard, so $A^* + D^*$ is smaller than $A + D$. If, for example, $a > c$, which means that as many true masters as possible would be preferred, then

$$(A^* + D^*)a + (C^* + B^*)\,c < (A + D)a + (C + B)c. \quad (9)$$

From this result it is clear that the expected utility on the second occasion might be smaller than the expected utility on the first occasion. Clearly, a loss through the use of optimal cutoff scores has been obtained because it affected in an adverse way the amount of learning that took place. Admittedly, there is limited empirical evidence presently for the hypothesized interaction between student knowledge of cutoff scores and amount of learning, but the hypothesis seems plausible, and it is supported by results in a study by Block (1972).

The Cutoff Score as a Target for Examinees

Some educators have argued that students ought to know the standard by which their achievements will be evaluated. In fact, students are often informed of the standard that they are expected to meet. Knowing the standard on the domain score scale, π_0, will not mean much to students if the cutoff score on the actual test differs substantially. In those situations where students are informed of the standard expected of them, optimal cutoff scores differing from the standard cannot be applied. It should be remembered that the optimal cutoff score is based not only upon the standard but also upon the population distribution, and this population distribution may not mean very much to any particular examinee. Students would be quite annoyed if the optimal cutoff score was higher than they were told ($n\pi_0$). Also, in the context of, say, basic skills programs, judges who labored hard to set a reasonable standard, as well as the general public, might have difficulties when there is a discrepancy between the standard and the cutoff score.

In some curricula, examinees have obtained the right to be informed beforehand about the value of the cutoff score; the cutoff score is thus fixed beforehand. When the cutoff score is determined beforehand, it seems that the test administrator is left

without a tool to classify examinees in an optimal way. Should it be concluded, then, that decision theory has nothing to tell the test administrator? Such a conclusion would perhaps be premature. Rather, the decision problem should be formulated in another way. A cutoff score should be found that *if known beforehand*, would lead to a maximal expected utility. Clearly, this problem is more difficult than the problem of finding the optimal cutoff score after the test results have been obtained. In fact, the approach needs some experimentation with cutoff scores. Initially examinees could be passed with an observed proportion correct equal to or exceeding π_0, for lack of an alternative. This means that C_x should be set equal to the first integer equal to or exceeding $n\pi_0$. This procedure has a disadvantage, however, because the observed score distribution is discontinuous. A procedure that alleviates this problem has been proposed by Wilcox (1976).

Conclusions

The substantial problems described in this paper may suggest that decision theory should be discarded as an approach for setting cutoff scores. Actually, decision theory as such provides the correct framework for decision making, but the problems described in this paper will need to be overcome and/or understood more thoroughly before the full potential of the approach can be realized. Studies and developments along the lines of those suggested here should substantially improve the situation and provide a more adequate basis for determining the worth of decision theory and how that theory might be successfully applied to the problem of setting cutoff scores.

References

Block, J. H. (1972). Student learning and the setting of mastery performance standards. *Educational Horizons*, *50*, 183–190.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277–289.

De Gruijter, D. N. M. (1980, June). *Accounting for the uncertainty in performance standards*. Paper presented at the International Symposium on Education Testing, Antwerp, Belgium (ERIC No. ED 199 280).

Educational Development Corporation. (1980). *Individualized criterion-referenced tests technical manual*. Tulsa OK: Author.

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

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

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

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*, 1–47.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika*, *41*, 65–78.

Isaacs, G. L., & Novick, M. R. (1978). Computer-assisted data analysis—1978. Manual for the computer-assisted data analysis (CADA) Monitor. Iowa City IA: The University of Iowa.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, *16*, 247–260.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239–269.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Maritz, J. S. (1966). Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika*, *53*, 417–429.

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

Mellenbergh, G. J., & Van der Linden, W. J. (1981). The linear utility model for optimal selection. *Psychometrika*, *46*, 283–293.

Novick, M. R. (1980). Statistics as psychometrics. *Psychometrika*, *45*, 411–424.

Novick, M. R., & Ellis, D. D. (1977). Equal opportunity in educational and employment selection. *American Psychologist*, *32*, 306–320.

Novick, M. R., Isaacs, G. L., & Dekeyrel, D. F. (1977). Computer-assisted data analysis—1977. Manual for

the computer-assisted data analysis (CADA) Monitor. Iowa City IA: Iowa Testing Programs.

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

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15*, 297–300.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education*. Washington DC: National Council on Measurement in Education.

Van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4*, 469–492.

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

Vijn, P., & Molenaar, I. W. (1981). Robustness regions for dichotomous decisions. *Journal of Educational Statistics, 6*, 205–235.

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

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics, 2*, 289–307.

Acknowledgment

The paper was prepared while the second author was on leave from the University of Massachusetts, Amherst. Without implying his support for the arguments in the paper, the authors are grateful to Wim J. van der Linden for his constructive criticism of an earlier draft.

Authors' Addresses

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leyden, Boerhaavelaan 2, 2334 EN Leyden, The Netherlands; or Ronald K. Hambleton, University of Massachusetts, Laboratory of Psychometric and Evaluative Research, Hills South, Room 152, Amherst MA 01003, U.S.A.