ED 199 280 TM 810 189

AUTHOR deGruijter, Dato N. M.
TITLE Accounting for the Uncertainty in Performance Standards.
PUB DATE Jun 80
NOTE 16p.: Paper presented at the International Symposium on Educational Testing (4th, Antwerp, Belgium, June 24-27, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores: Difficulty Level; Error of Measurement: *Mastery Tests: *Mathematical Models: Sampling: *Test Interpretation: Test Items: True Scores
IDENTIFIERS *Uncertainty

ABSTRACT

The setting of standards involves subjective value judgments. The inherent arbitrariness of specific standards has been severely criticized by Glass. His antagonists agree that standard setting is a judgmental task but they have pointed out that arbitrariness in the positive sense of serious judgmental decisions is unavoidable. Further, small misplacements of the standard therefore can be considered inconsequential. In this paper, the uncertainty with respect to the 'true' standard is quantified and the consequences of the specification of the uncertainty on the optimum passing score are studied. In a second approach the assumption is made that the standard setters not only have information with respect to the position of the standard, but also relative information with respect to the level of the target group of examinees. This information can be used in the final setting of the standard. If performance is lower than expected, one does better by lowering the standard a certain amount by means of a preconceived strategy. (Author/RL)

ACCOUNTING FOR THE UNCERTAINTY IN PERFORMANCE STANDARDS

Dato N.M. de Gruijter, University of Leyden

Abstract

The setting of standards involves subjective value judgments. The inherent arbitrariness of specific standards has been severely criticized by Glass in a special issue on standard setting in the 1978 volume of the Journal of Educational Measurement. His antagonists agree that standard setting is a judgmental task but they have pointed out that arbitrariness in the positive sense of serious judgmental decisions is unavoidable. Further, small misplacements of the standard therefore can be considered inconsequential.

The point of view in this paper is that the argumentation should not remain on the verbal level only, i.e. it is proposed to quantify the uncertainty with respect to the 'true' standard and to study the consequences of the specification of the uncertainty on the optimum passing score.

In a second approach the reasonable assumption is made that the standard setters not only have information with respect to the position of the standard, but also relative information, i.e. information with respect to the level of the target group of examinees. This information can be used in the final setting of the standard. If e.g. performance is lower than expected, one does better by lowering the standard a certain amount by means of a preconceived strategy.

Accounting for uncertainty in performance standards*

Dato N.M. de Gruijter
Educational Research Center
University of Leyden

Introduction

In the decision-theoretic approach to mastery decisions utilities or losses for pasing and failing are defined as functions of domain scores. For each examinee an observed score is obtained and that decision which results in the smaller expected loss is made.

The utility or loss structure uses - explicitly or implicitly - the concept of a standard of performance. In threshold loss e.g., the loss associated with failing an examinee is larger than zero for domain scores $\pi \geq \pi_0$, where π_0 is the performance standard, and equals zero for $\pi < \pi_0$. In more realistic continuous utility functions one might call the break-even value, the value of π for which the utilities for passing and failing are equal, the standard.

In the early decision-theoretic literature on mastery testing the standard is taken for granted. In more recent years one sees attempts by psychometricians to develop methods specifying utilities, and so implicitly standards, more carefully (see Novick and Lindley, 1978). The more advanced techniques, however, require subjective values judgments, which introduce a source of mistakes. This may become apparent when experts disagree with respect to the specification of utilities.

Glass (1978), in a special issue on standard setting in the Journal of Educational Measurement, severely criticizes the standard setting approach in testing, referring to the arbitrariness of standards; and some contemporary practices may serve as examples demonstrating that he is not entirely wrong. His opponents (e.g. Popham, 1978) agree that standard setting is a judgmental task, but they do point out that arbitrariness, in the positive sense of serious judgmental decision making, is unavoidable. Block (1978), referring to Schwab (1969), suggests in connection with this discussion that one cannot expect right solutions, but at best defensible solutions.

* Paper presented at the Fourth International Symposium on Educational Testing, Antwerp, June 24-27, 1980.

Moreover, Scriven (1978) reminds us that there is a range of domain scores for which passing or failing does not make much difference; this is reflected by realistic utility functions. Therefore small misplacements of the standard may be considered inconsequential.

Although uncertainty with respect to the adequacy of a given standard is generally admitted, it has not been given a formal treatment within the decision-theoretic approach to mastery testing. One of the aims of this paper is to provide such a treatment in connection with threshold loss.

Part of the uncertainty with respect to the most adequate position of the standard may be due to the fact that a readily available type of information, normative information, is often neglected. The use of normative information seems incompatible with the philosophy of mastery testing. It will be argued, however, that in fact it is not. How normative information may be used in standard setting, will be demonstrated.

Threshold loss and uncertainty in standard setting

Let us assume that a standard π_0 has been set. Further, let us assume that threshold loss represents an adequate approximation to losses due to classification errors, i.e. the following loss structure is used:

	$\pi < \pi_0$	$\pi \geq \pi_0$
pass	L_{01}	0
fail	0	L_{10}

with L_{01}, $L_{10} > 0$. In fact, one only needs to determine the loss ratio $L = L_{01}/L_{10}$. Therefore, in the following L_{01} is replaced by L, and L_{10} by 1. Let us further assume that each examinee answers the same number of items n and that the binomial error model holds. This means that the observed score x is a sufficient statistic for π.

Examinee p should be passed if the expected loss on failing (the probability of mastery times the loss due to failing a master) exceeds the expected loss on passing.

(1) $$\mathcal{E}\{\pi_p \geq \pi_0 | x_p\} > L\mathcal{E}\{\pi_p < \pi_0 | x_p\}.$$

The opposite action is taken in case the inequality sign is reversed, while we are indifferent with respect to the action taken in case of an equality sign. The conditional probabilities in (1) can be computed by the application of Bayes' theorem in case the distribution of π is known. The first score x for which (1) is satisfied, is the optimal cutting score.

In the binomial error model the error variance varies as a function of π, a troublesome characteristic in some analyses. Therefore, in this paper transformed domain scores $\gamma = \sin^{-1}\sqrt{\pi}$ are used instead: this inverse sine transformation is a variance-stabilizing transformation (see e.g. Novick et al., 1973). For observed proportions the following transformation is chosen:

$$g_p = \sin^{-1}\sqrt{(x_p + 3/8)(n + 3/4)}.$$

The distribution of g_p can be approximated by $N(\gamma_p, (4n + 2)^{-1})$, which implies that the error variance on the transformed scale is independent of γ.

Assuming that γ is approximately $N(\mu_\gamma, \phi_\gamma)$ and the sample size is large - which means that μ_γ and ϕ_γ are accurately estimated - the posterior distribution of the transformed domain score of examinee p, γ_p, is approximately normal with a mean equal to

$$\text{(2)} \qquad \hat{\gamma}_p = \rho g_p + (1-\rho)\,\mu_\gamma$$

where

$$\text{(3)} \qquad \rho = \frac{\phi_\gamma}{\phi_\gamma + (4n+2)^{-1}}$$

and a variance equal to

$$\text{(4)} \qquad \phi = \rho(4n+2)^{-1}.$$

Equation (2) is a Kelley-estimate of examinee's p transformed domain score. Using inequality (1) the criterion for passing becomes

$$\text{(5)} \qquad 1 - \Phi[\phi^{-\frac{1}{2}}(\gamma_0 - \hat{\gamma}_p)] \quad 1.\Phi[\phi^{-\frac{1}{2}}(\gamma_0 - \hat{\gamma}_p)]$$

where Φ is the cumulative normal distribution and $\gamma_0 = \sin^{-1} \sqrt{\pi_0}$ (cf. Equation 1).

Now we are ready to introduce uncertainty with respect to π_0 and, for that matter, γ_0. Let us assume that this uncertainty may be reflected by a normal distribution for γ_0, $N(\mu_{\gamma_0}, \phi_0)$.

Here γ_0 may be the standard agreed upon and ϕ_0 is determined in such a way that it reflects the amount of uncertainty with respect to the correct value of γ_0. The distribution of $\gamma_p - \gamma_0$ determines the optimal decision (passing or failing). One obtains for this distribution

$$(6) \qquad \gamma_p - \gamma_0 \overset{\sim}{-} N(\bar{\gamma}_p - \mu_{\gamma_0}, \phi + \phi_0).$$

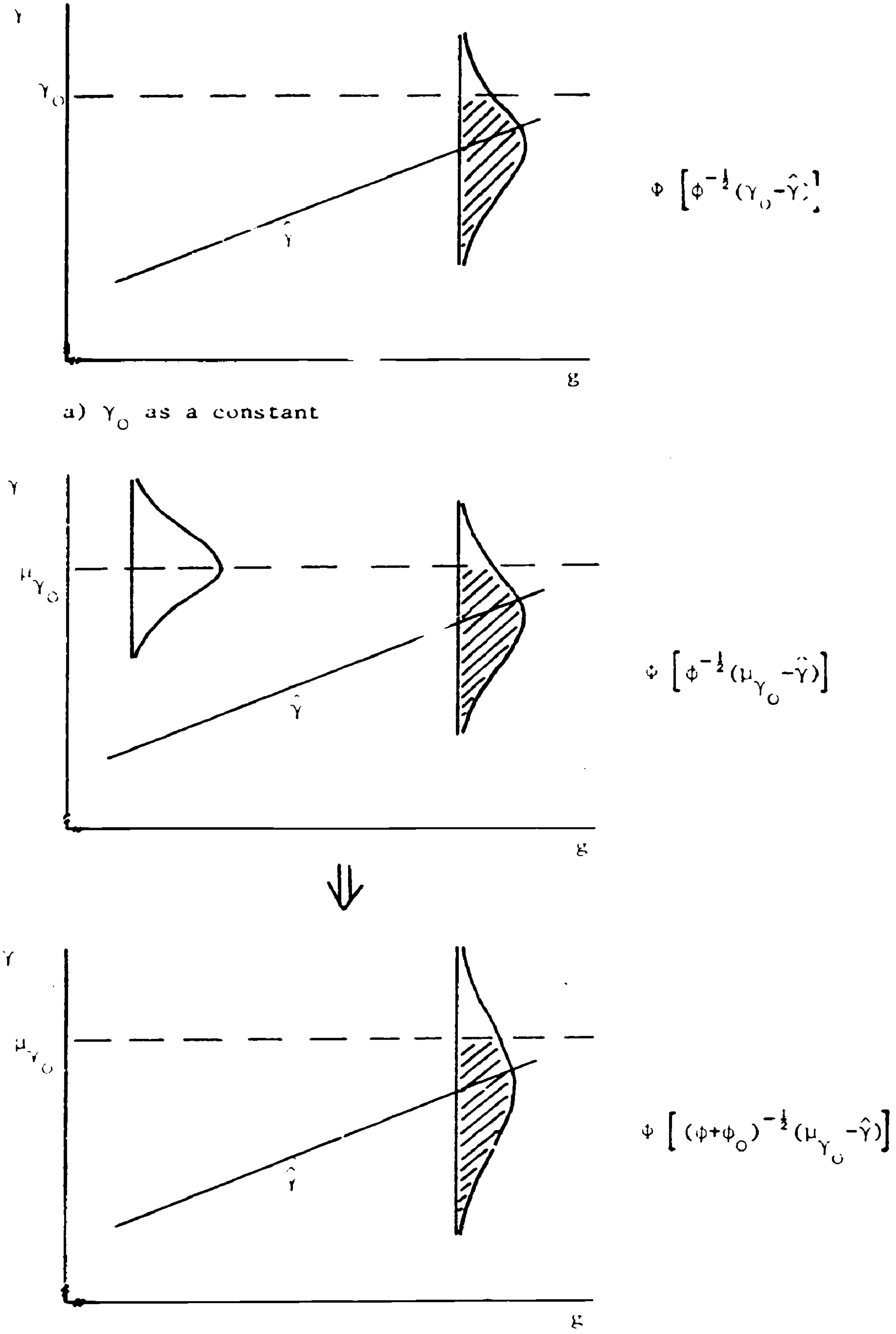
The criterion for passing becomes

$$(7) \quad 1 - \Phi[(\phi+\phi_0)^{-\frac{1}{2}}\{0-(\hat{\gamma}-\mu_{\gamma_0})\}] > L\Phi[(\phi+\phi_0)^{-\frac{1}{2}}\{0-(\hat{\gamma}-\mu_{\gamma_0})\}].$$

Of course one has to be aware of the fact that in the above derivation of the distribution of $\gamma_p - \gamma_0$ errors in the estimation of $\gamma_p - \gamma_0$ for different persons p are correlated due to the common variation in γ_0. In Figure 1 both cases, the case of variable γ_0 and the case of fixed γ_0, are displayed.

For a loss ratio equal to one decisions are not affected by the introduction of uncertainty in γ_0: one obtains indifference between the decision to pass and the decision to fail for $\hat{\gamma}_p = \mu_{\gamma_0}$ (or $\hat{\gamma}_p = \gamma_0$). For other loss ratios, however, tne introduction of uncertainty in γ_0 may make a difference.

Making optimal decisions by the minimization of expected loss under uncertainty with respect to γ_0, is only one side of the problem. It is also important to study the effects of different possible values of γ_0 on decision making. This calls for a robustness study (Vijn, 1980) in which intervals for γ_0 characterized by the same optimal cutting score are computed. In fact, it would be even better to study the joint effect of variation in γ_0 and L on the optimal cutting score since in this way one also makes allowance for uncertainty with respect to the proper choice of L.


γ
γ_0
$\hat{\gamma}$
g
$\Phi\left[\phi^{-\frac{1}{2}}(\gamma_0-\hat{\gamma})\right]$
γ
μ_{γ_0}
$\hat{\gamma}$
g
$\Phi\left[\phi^{-\frac{1}{2}}(\mu_{\gamma_0}-\hat{\gamma})\right]$
γ
μ_{γ_0}
$\hat{\gamma}$
g
$\Phi\left[(\phi+\phi_0)^{-\frac{1}{2}}(\mu_{\gamma_0}-\hat{\gamma})\right]$

a) γ_0 as a constant

b) γ_0 as a variable

Figure 1. Posterior probability $P(\gamma \geq \gamma_0)$ in two cases

This results in regions in the two-dimensional (γ_0, L) space corresponding to particular values for the cutting score.

The idea to vary L is not a new one in the decision theoretic approach to mastery testing. Some authors have presented their data in such a way that the consequences of values of L, other than the one preferred by these authors, can be determined. This means that the reader who prefers another loss ratio, may examine the results from his or her own point of view. Furthermore, such a presentation is useful in case of uncertainty with respect to L, as has been suggested above. A nice example of such a presentation is an article by Huynh (1977).

In the following example of a robustness study in which γ_0 and L are varied, I shall use Mellenbergh _et al_.'s (1977) data from 184 examinees on a 19 item mastery test. Their data can be fitted by a normal distribution for g with mean

$$\bar{g} = \hat{\mu}_\gamma = 1.128$$

and variance

$$s_g^2 = \hat{\sigma}_g^2 = .024.$$

The variance of γ equals ϕ_γ = .011, the posterior variance of γ equals ϕ = .006 (estimates are assumed to equal the true values here).

Using a loss ratio equal to 2 and a standard γ_0 equal to 1.107 (corresponding to π_0 = .80), the optimal cutting score equals g = 1.217 (corresponding to an observed score of 17). The same optimal cutting score is obtained if the fixed γ_0 is replaced by the distribution N(1.107, .0025). The results were obtained replacing the cumulative normal distribution by the closely related cumulative logistic distribution

$$\exp(1.7t)/\{1 + \exp(1.7t)\} \simeq \Phi(t)^{**}$$

** I have taken the scale factor 1.7 which generally is used while using this scale factor, the cumulative logistic differs by less than 0.01 from the cumulative standard normal distribution for all values t. Molenaar (1974) demonstrates that the scale factor 1.6 brings the densities in close agreement while for $|t| < 1$ the agreement between the cumulative distributions is close.

which means that for fixed γ_0 indifference between passing and failing obtaines if

$$(8a) \qquad 1.7\ \phi^{-\frac{1}{2}}\ (\gamma - \gamma_0) = \log L$$

and for γ_0 as a variable it

$$(8b) \qquad 1.7\ (\phi+\phi_0)^{\frac{1}{2}}\ (\hat{\gamma}-\mu_{\gamma_0}) = \log L.$$

From (8a) it is easy to obtain boundaries of indifference regions by substituting $\hat{\gamma}$ from (2) for several values of g(x). For a particular value of x Equation (8a), with log L as a function of γ_0, defines the boundary between the region where x is the optimal cutting score and the region where x+1 is the optimal cutting score; one may verify this by substituting a 'greater than' sign in Equation (8a) for passing. In Figure 2 regions are given for $1 \leq L \leq 3$ and $\mu_{\gamma_0} - 2\phi_0 \leq \mu_{\gamma_0} + 2\phi_0$.

Conspicuous is the sensitiveness of the optimal cutting score to changes in γ_0 due to the unreliability of the test, i.e. the strong regression of $\hat{\gamma}_p$ to μ_γ.

Since the result is disappointing, one should put a lot of effort in diminishing the uncertainty with respect to γ_0. For example, supposing that only one expert has been used in setting the standard and that ϕ_0 reflects the uncertainty with respect to the resulting standard, one may reduce the uncertainty by having the standard set by more than one expert. The underlying critical assumption is that expert opinion is unbiased (Jaeger, 1979).

One may wonder what would have happened if a more realistic loss function than threshold had been used instead. Take for instance linear loss with respect to failing

$$L_{fail}\ (\gamma) = 0$$
$$L_{pass}\ (\gamma) = -a\ (\gamma - \gamma_0) \qquad a > 0.$$

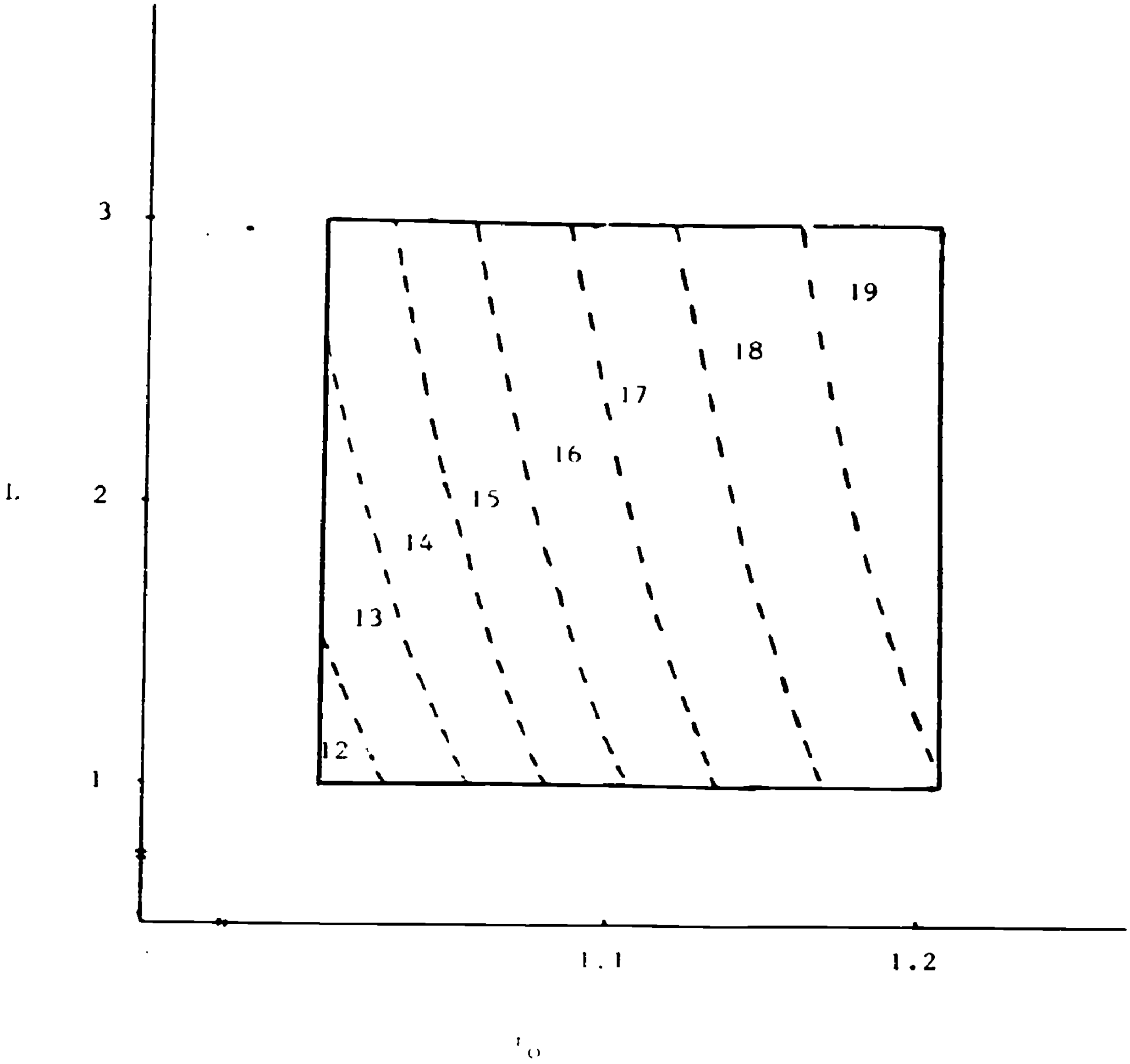


Figure 2. Robustness regions for r_o and L. Numbers in the figure give optimal cutting scores.

Here the optimal cutting score is $\dot{\gamma} = \gamma_0$ and a change $\Delta\gamma_0$ corresponds to a change $(1/\rho)\Delta\gamma_0$ in the optimal cutting score g (treating g as a continuous variable). With respect to robustness to changes in γ_0 there is no difference with threshold loss. However, there might be a different interpretation of the seriousness of a change in the cutting score due to the characteristics of the linear loss function, which weights wrong classifications of borderline examinees less heavily.

The use of normative information in standard setting

In the previous section I have demonstrated that, if there is a range of plausible γ_0 values, there are several alternative cutting scores. Here I will try to demonstrate that normative information may be used in standard setting; the use of extra information reduces the range of plausible γ_0 values. A proposal based on another philosophy has been presented by Hofstee in this symposium.

Clearly both instruction and examination standards should be geared to the level of the target group of examinees, i.e. their entrance characteristics in a broad sense. If performance with respect to a given standard is relatively low, one may enlarge the percentage of masters by lengthening the instructional period, but this is only done with diminishing returns. In such a situation, lowering the standard could be the more realistic action.

In this way the actual examinee results play a role in standard setting. One could say with Hofstee (1973) that realistic standards in the end are normative A similar point of view is taken by Shepard (1979).

At first sight normative ideas look incompatible with the philosophy of criterion-referenced testing. The use of normative information, however, does not mean the introduction of an arbitrary group of persons and an arbitrary passing percentage defined for this group. Neither does it mean that the standard varies with the group of examinees, a procedure which correctly is rejected in criterion-referenced measurement. Here only the claim is made that normative information can be useful in reducing uncertainty in standard setting.

In practice it may be difficult to make a distinction between normative and absolute information of performance. For instance, information on performance in comparable or subsequent instructional programs can be useful in standard setting. But the evaluation of that performance depends on arbitrary decisions with respect to the standard in those programs, possibly based on a mixture of normative and absolute arguments. For this reason Glass (1978) criticized Huynh's (1976) referral loss as bootstrapping on other criterion scores. Nevertheless, more or less vague ideas concerning the percentage of examinees who have a satisfactory performance exist; it is on this basis that Shepard (1979) suggests the correction of the standard if the percentage of failures seems inordinate. In my opinion such feelings should be formalized beforehand, so that we may state a priori for all possible outcomes on an examination which decision should be made. Such a procedure is more satisfactory than a simple correction by hindsight.

Assume that the vague normative knowledge can be formalized in a density. Specifically let $F' = \log(F/(1-F)$, where F is the proportion of examinees thought to be satisfactory, be $N(\mu_{F'}, \phi_{F'})$; F itself has approximately a beta distribution.

Further, let γ_0 be $N(\mu_{\gamma_0}, \phi_0)$ as in the previous section and let us assume that γ_0 and F' are independently distributed. Equiprobability contours of the bivariate distribution of γ_0 and F' are given by

$$(9) \qquad \phi_0^{-1}(\gamma_0 - \mu_\gamma)^2 + \phi_{F'}^{-1}(F' - \mu_{F'})^2 = \text{constant}.$$

Assuming that the cumulative distribution of γ can be approximated by the cumulative logistic, we obtain the following relation between transformed domain score and transformed proportion of examinees exceeding the transformed domain score

$$(10) \qquad F' = -1.7\,\phi_\gamma^{-\frac{1}{2}}(\gamma - \mu_\gamma).$$

The line defined by (10) is tangent to one of the ellipses defined by (9). The point of contact defines values γ_0 and F' satisfying (10) having the highest joint probability of occurance; this value of γ_0 is the new standard.

It is easy to deduce that points of contact for equations (10) differing only in the value of μ_γ, are lying on a line through $(\mu_\gamma, \mu_{F'})$ with a slope equal to $(\phi_\gamma^{\frac{1}{2}}/1.7)(\phi_{F'}/\phi_0)$.

In Figure 3 the procedure is demonstrated for the Mellenbergh et al. data. Here $\mu_0 = 1.107$, $\mu_{F'} = .847$ (corresponding to a proportion F equal to .70) and $\phi_{F'}/\phi_0 = 100$. In the figure the line is drawn connecting optimal points (F', γ_0) for distributions having the same variance on the transformed scale as the Mellenbergh *et al.* data. The intersection of the two lines in the figure gives the corrected standard ($\gamma_0 = 1.083$). For this standard the optimal cutting score, given L = 2, is 16 instead of 17. Figure 4 gives the relationship between F and π_0 for distributions having the same ϕ_γ as in the example; the relation between F and π_0 in this case is surprisingly linear in the range of interest.

One may choose the joint distribution of γ_0 and F' in such a way that for a specific distribution of γ predetermined values of γ_0 and F' are obtained. This is very useful when a new program is to replace an old one. In case the unfortunate outcome is that the new program did not result in a learning change, one supposedly would like to stick to the old value of γ_0. If, however, the mean score remains the same, but the variance changes, it is not possible to have the same γ_0 using this procedure. If one would like to keep the old standard in this case, another procedure is in order; such a proposal has been made earlier (De Gruijter, 1978) using prior information with respect to the population mean instead of prior information with respect to F'.

Two remarks remain to be made. First, I have assumed that the population distribution of γ is known while in practice only sample information is available and sometimes only after more testing occasions population parameters can be accurately estimated by combining the data. Secondly, in the proposal it is assumed that information with respect to the proportions of satisfactory examinees is available, while Shepard defines the problem in terms of proportion passing. The use of proportion satisfactory is preferable while proportion passing also depends on test length.

Examinations varying in difficulty level

The binomial error model can be applied in case every examinee has to answer a different set of items randomly chosen from the item domain or in case items have approximately the same difficulty level. Otherwise, the compound binomial error model may be appropriate.

The use of tests composed of items varying in difficulty level, presents some complications: if items vary in difficulty level, so will the tests. This means that the standard appropriate for a particular test is defined

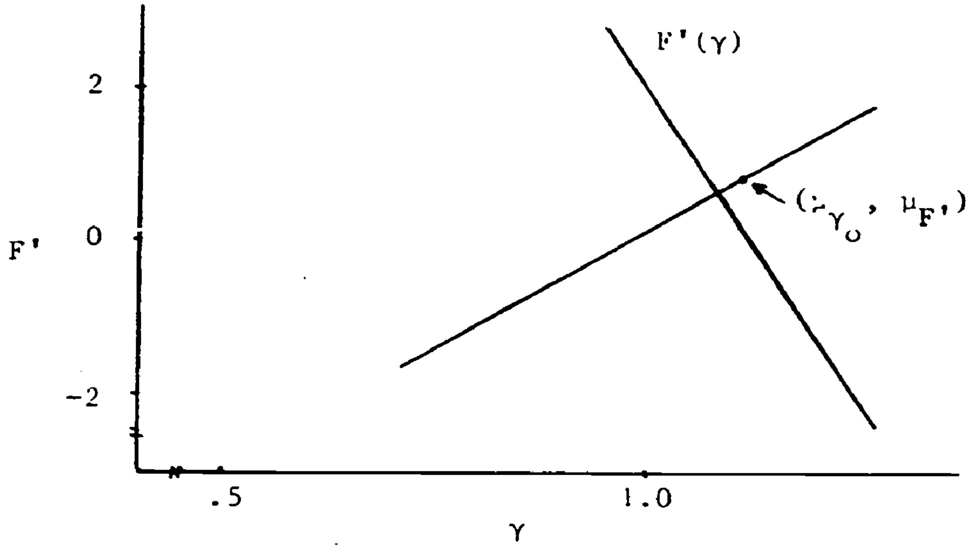


Figure 3. Transformed percentage satisfactory as a function of γ and the line defining points of contacts for distributions with identical ϕ_γ.

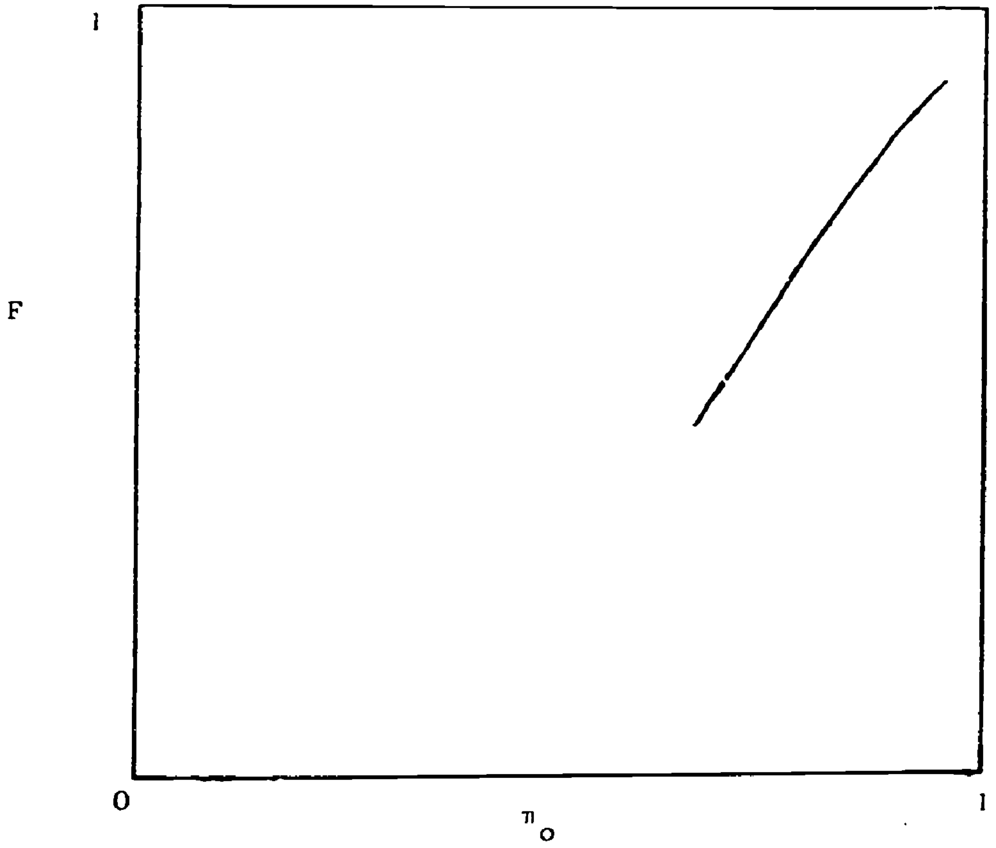


Figure 4. The relationship between π_o and F for population distributions with identical ϕ_γ.

in terms of the relative true score scale of this test and that this standard may differ to an unknown extent from the standard on the domain score scale - where the domain score is the expected relative true score with expectation taken over all possible test forms - and the standards on other test forms.

Interestingly enough, if the groups of examinees taking different test forms are all large and may be considered random samples from the population of examinees, standards should be transferred from one test form to another by a relative procedure. This again is an example where a relative approach fits into the framework of criterion-referenced measurement (De Gruijter, 1978). A critical assumption is the randomness assumption; it does not hold if people react to the mean level of the group of persons with which they study. However, the coherence of large groups (probably consisting of many small groups) is small while, further, mean levels of different large groups do not differ much in case there are no systematic factors effecting differences.

Summary

It is argued that a full decision-theoretic approach to criterion-referenced measurement should incorporate uncertainty with respect to the standard. Furthermore, it is demonstrated that normative information in itself is not incompatible with the idea of criterion-referenced measurement. On the contrary, normative information may be used in order to determine more precisely a satisfactory standard.

References

Block, J.H. Standards and criteria: a response. *Journal of Educational Measurement*, 1978, *15*, 291-295.

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

Gruijter, D.N.M. de. A Bayesian approach to the passing score problem. *Tijdschrift voor Onderwijsresearch*, 1978, *3*, 145-151.

Hofstee, W.K.B. Een alternatief voor normhandhaving bij toetsen. *Nederlands Tijdschrift voor de Psychologie*, 1973, *28*, 215-227.

Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, *41*, 65-78.

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. *Psychometrika*, 1977, *42*, 601-608.

Jaeger, R.M. Measurement consequences of selected standard-setting models. In M.A. Bunda and J.R. Sanders (Eds): *Practices and problems in competency-based education*. Washington: National Council on Measurement in Education, 1979.

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

Molenaar, W. De logistische en de normale kromme. *Nederlands Tijdschrift voor de Psychologie*, 1974, *29*, 415-420.

Novick, M.R., Lewis, C. and Jackson, P.H. The estimation of proportions in *m* groups. *Psychometrika*, 1973, *38*, 19-46.

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

Schwab, J.J. The practical: a language for curriculum. *School Review*, 1969, *78*, 1-24.

Scriven, M. How to anchor standards. *Journal of Educational Measurement*, 1978, *15*, 273-275.

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

Vijn, P. *Prior information in linear models*. Unpubl. Ph. Thesis, Univ. Groningen, 1980.