A Note on Decision Theoretic Coefficients for Tests

Rand R. Wilcox
University of California at Los Angeles

Recently it was suggested that the Bayes risk might be used to characterize tests. To conform to common practices about indexes, a rescaling of the Bayes risk was proposed. The motivation for this new coefficient, δ, was to provide an index that has a large value when the Bayes risk is small and that has a value in the closed interval [0, 1]. However, since δ might have a value outside this range, a modification of δ is described which yields an index that always has a value between zero and one.

Recently, van der Linden and Mellenbergh (1978) argued in favor of using decision theoretic techniques for characterizing a test; they pointed out that the Bayes risk is a natural index to use from this point of view. Van der Linden and Mellenbergh also noted that the Bayes risk has two disadvantages. First, it is conventional in test theory to define indexes so that the scale has a direction opposite to that in which the Bayes risk is represented. Second, although in test theory indexes are nearly always defined on the closed interval [0, 1], the range of the possible values of the Bayes risk can be different.

Let $L(\hat{\theta}(x)|\theta)$ be the loss function associated with a particular decision problem which is a function of the decision rule, $\hat{\theta}(x)$; the examinee's true score, θ; and the examinee's observed score, x. For a particular examinee, the risk or expected loss is $R = EL(\hat{\theta}(x)|\theta)$. The Bayes risk is $R_B = E_\theta R$, where E_θ means expectation with respect to the probability density of θ.

In an attempt to correct the two deficiencies described above, van der Linden and Mellenbergh suggest the index

$$\delta = (R_n - R_B)/(R_n - R_c) , \qquad [1]$$

where R_c and R_n are the Bayes risks in the situations in which the test contains, respectively, complete and no information about the true scores. They point out, however, that the index δ does not correct the second difficulty, since δ might not be in the interval [0, 1]. The purpose of this note is to indicate a simple modification of δ that solves this problem.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 4 Fall 1978 pp. 609-613
© Copyright 1978 West Publishing Co.

Coefficient γ

To ensure that δ has a value between 0 and 1, R_n is replaced with R_u, the least upper bound on R_B; and R_c is replaced with R_t, the greatest lower bound on R_B. It is assumed that the decision rule, $\hat{\theta}(x)$, has been specified and that there is an arbitrary joint density for x and θ. In this case δ becomes

$$\gamma = (R_u - R_B)/(R_u - R_t) \quad . \qquad [2]$$

By requiring R_u to be the least upper bound on R_B, rather than just any upper bound, it is guaranteed that $\gamma \geqslant 0$. By also choosing R_t in the manner described above, it is guaranteed that $\gamma \leqslant 1$. It is evident that $0 \leqslant \gamma \leqslant 1$, as was required. Note that if R_B is defined on an open interval, $0 < \gamma < 1$. For many practical situations R_u and R_t can be evaluated. This point can be illustrated by re-examining the situations considered by van der Linden and Mellenbergh.

Point Estimation

Suppose the purpose of a test is to estimate true score, that $\hat{\theta}$(x) is a decision rule for estimating θ based on the observed score, x, and that squared error loss is used, i.e., $L = (\hat{\theta}(x) - \theta)^2$. If, as is usually the case, the observed scores can have the values 0,1, . . . ,n (n being the number of items on the test), the Bayes risk is given by

$$R_B = \sum_{x=0}^{n} \int_0^1 (\hat{\theta}(x) - \theta)^2 \, f(x|\theta) g(\theta) d\theta \ , \qquad [3]$$

where $f(x|\theta)$ is the conditional probability function of observed scores and
$g(\theta)$ is the density function of true scores.

The least upper bound of R_B depends on such things as the assumptions made in a particular testing model, the statistic $\hat{\theta}(x)$, and the number of items in the test. A situation in which a minimum number of assumptions are made can be considered first. Then two specific forms for $\hat{\theta}(x)$ can be considered.

For notational convenience it is assumed that $0 \leqslant \theta \leqslant 1$. Since true score is typically defined as an expected value, it is usually possible to multiply this expected value by an appropriately chosen constant so that θ will be in this range. This rescaling has no effect on the value of γ. Since $0 \leqslant \theta \leqslant 1$, it makes little sense to allow $\hat{\theta}(x) < 0$ or $\hat{\theta}(x) > 1$; therefore, it is also assumed that $0 \leqslant \hat{\theta}(x) \leqslant 1$.

Except for highly unusual circumstances, it will be the case that $\hat{\theta}(x)$ is a monotonically increasing function of x. It follows that the least upper bound of $(\hat{\theta}(x) - \theta)^2$ is either $(\hat{\theta}(0) - 1)^2$ or $(\hat{\theta}(n) - 0)^2$, whichever is largest. Let L_2 denote the larger of these two quantities. Then $R_B \leqslant L_2$, since L_2 is an upper bound to $(\hat{\theta}(x) - \theta)^2$. For the case $L_2 = (\hat{\theta}(0) - 1)^2$, $R_B = L_2$ when $Pr(x{=}0, \theta{=}1) = 1$. If $L_2 = (\hat{\theta}(n) - 0)^2$, $R_B = L_2$ when $Pr(x{=}n, \theta{=}0) = 1$. Note that this least upper bound was derived under the assumption that any joint probability density function for θ and x is possible. If it is assumed that $E(\hat{\theta}(x)|\theta) = \theta$, then in general L_2 will not be the least upper bound. Such a case is considered below.

The greatest lower bound for R_B is zero. This occurs when $Pr(\theta{=}\hat{\theta}(x)) = 1$ for all x and θ (i.e., perfect estimation). Thus, γ may be written as

$$\gamma = (L_2 - R_B)/L_2 . \qquad [4]$$

As previously indicated, the derivation of L_2 was made without any restriction on the joint probability density function of x and θ or the decision rule $\hat{\theta}(x)$. In general, if it is assumed that $E(\hat{\theta}(x)|\theta) =$

θ, L_2 is no longer the least upper bound to R_B. Accordingly, a least upper bound to R_B is derived when an unbiased estimate of θ is used.

Consider the case in which $x = 0$ or 1. Without making any assumption about the form of $h(x|\theta)$, it follows that

$$E[(x-\theta)^2|\theta] \leq \tfrac{1}{4} \quad , \tag{5}$$

with equality holding when $Pr(x{=}0) = Pr(x{=}1) = \tfrac{1}{2}$ (e.g., see Skibinsky, 1977). Taking the expectation of both sides of Equation 5,

$$E_\theta[(x-\theta)^2] \leq \tfrac{1}{4}. \tag{6}$$

Now suppose that for an examinee, there are n observations $x_1 \ldots, x_n$ with x_i independent of x_j, $i \neq j$, $x_i = 0$ or 1 and that $E(x_i|\theta) = \theta$, $i = 1, \ldots, n$. Thus,

$$\bar{x} = n^{-1} \sum_i x_i \tag{7}$$

is an unbiased estimate of θ, and

$$E[(\bar{x}-\theta)^2|\theta] \leq 1/4n \quad , \tag{8}$$

with equality holding when $Pr(\hat{\theta}(x_i){=}0) = Pr(\hat{\theta}(x_i){=}1) = \tfrac{1}{2}$, $i = 1, \ldots, n$. It follows that $1/4n$ is the least upper bound on R_B. The greatest lower bound is zero, which means that $\gamma = 1-4nR_B$. Griffin and Krutchkoff (1971) indicate that $R_B = \sigma_x^2 - \sigma_\theta^2$ where σ_x^2 is the variance of the marginal distribution of x and σ_θ^2 is the variance of true scores. Using results given by Lord and Novick (1968), σ_θ^2 can be estimated, which in turn yields an estimate of R_B and γ.

γ may not increase with increasing values of n. To reflect increased accuracy for larger values of n, R_u might be replaced with the least upper bound based on a single observation. In this case, γ becomes $\gamma = 1-4R_B$.

For the problem of point estimation, van der Linden and Mellenbergh concentrate on Kelly's linear regression estimate of true scores, namely

$$\hat{\theta}(x) = \rho_{xx'}\bar{x} + (1-\rho_{xx'})E(\bar{x}) \quad , \tag{9}$$

where $\varrho_{xx'}$ is the reliability of the test. Since x is an unbiased estimate of the examinee's true score and if it is assumed that the first two moments of both $h(x|\theta)$ and $g(\theta)$ exist, then from Griffin and Krutchkoff it follows that Equation 9 is optimal in the sense that it is the linear estimate minimizing the Bayes risk. An interesting theoretical result given by van der Linden and Mellenbergh is that

$$R_B = \sigma_\theta^2(1-\rho_{xx'}) \tag{10}$$

when Equation 9 is used to estimate θ, which implies that $\delta = \varrho_{xx}$.

In practice, however, $\varrho_{xx'}$ is unknown. If $\varrho_{xx'}$ is estimated with, say, $\hat{\varrho}_{xx'}$ and the results are substituted in Equation 9, an estimate of Kelly's regression estimate of true score is obtained. It should be noted, however, that in this case the Bayes risk is no longer given by Equation 10. In fact, the Bayes risk takes on a much more complicated form which, as pointed out by Griffin and Krutchkoff, cannot be evaluated theoretically. Thus, it is unclear how to estimate R_B, which means that neither δ or γ can be estimated.

Dichotomous Decisions

Consider the problem of determining whether an examinee's true score, θ, is above or below a known constant, θ_0. The decision $\theta \geqslant \theta_0$ is made if $x \geqslant x_0$, where x_0 is a specified passing score. This problem has been considered by several authors in the context of a mastery test (Harris, 1974; Novick & Lewis, 1974; Huynh, 1976; Wilcox, 1977). Let ℓ_{10} be the loss incurred when $x \geqslant x_0$ and $\theta < \theta_0$. Let ℓ_{01} be the loss when $x < x_0$ and $\theta \geqslant \theta_0$. If a correct decision is made, it is assumed that the loss is zero. For a randomly selected examinee, the probabilities p_{ij} (i=0, 1; j=0, 1) associated with the four possible outcomes are shown in Table 1.

Table 1
Probabilities Associated With The
Four Possible Outcomes

		$x < x_0$	$x \geq x_0$
True Score	$\theta < \theta_0$	P_{00}	P_{01}
	$\theta \geq \theta_0$	P_{10}	P_{11}

Thus, p_{10} is the probability of $\theta \geqslant \theta_0$ and $x < x_0$. The Bayes risk is

$$R_B = \ell_{10}P_{10} + \ell_{01}P_{01}. \qquad [11]$$

As before, the least upper bound of R_B is highly dependent upon the situation at hand. Several cases can be considered.

First, note that it is always possible to guarantee that the probability of a correct decision for a given examinee is at least ½, simply by making the decision $\theta \geqslant \theta_0$ or $\theta < \theta_0$ at random. When this decision rule is used, the least upper bound of R_B is $R_u = ½\ell_{max}$, where $\ell_{max} = max[\ell_{10}, \ell_{01}]$. The greatest lower bound is $\ell_{min} = ½min[\ell_{10}, \ell_{01}]$ and so

$$\gamma = \frac{\ell_{max} - 2R_B}{\ell_{max} - \ell_{min}}. \qquad [12]$$

Next suppose that $f(x|\theta)$ is stochastically increasing. This means that $\theta < \theta'$ implies that $F(x|\theta) \geqslant F(x|\theta')$ for all x where $F(x|\theta)$ is the cumulative distribution corresponding to $f(x|\theta)$. The binomial and Poisson distributions are two examples where this is true. Observe that R_B may be written as

$$R_B = \ell_{01}F(x_o-1|\theta\geq\theta_o)Pr(\theta\geq\theta_o) + \ell_{10}[(1-F(x_o-1|\theta<\theta_o)]Pr(\theta<\theta_o). \qquad [13]$$

Thus, R_B is a weighted average of

$$\ell_{01}F(x_o-1|\theta\geq\theta_o) \qquad [14]$$

and

$$\ell_{10}[1-F(x_o-1|\theta<\theta_o)] \qquad [15]$$

and its value lies in the interval bounded by these two points.

Since $F(x|\theta)$ is stochastically increasing, the minimum of Equation 14 occurs at θ = 1, the minimum of Equation 15 occurs at $\theta = 0$, and the maximum of both occurs at $\theta = \theta_0$. It follows that

$$R_u = max[\ell_{01}F(x_o-1|\theta=\theta_o), \ell_{10}(1-F(x_o-1|\theta=\theta_o))] \qquad [16]$$

and

$$R_c = \min[\ell_{01}F(x_o-1|\theta=1),\ell_{10}(1-F(x_o-1|\theta=\theta_o))] \qquad [17]$$

As a final example, following Fhanér (1974) and Wilcox (in press), the problem of deciding whether θ is below θ_1 or above θ_2, $\theta_1 < \theta_2$ can be considered. If $\theta_1 < \theta < \theta_2$, either decision is said to be correct. The open interval (θ_1, θ_2) is called the indifference zone. In this case R_c is given by Equation 17 and

$$R_u = \max[\ell_{01}F(x_o-1|\theta=\theta_2),\ell_{10}(1-F(x_o-1|\theta=\theta_1))]. \qquad [18]$$

References

Fhanér, S. Item sampling and decision making in achievement testing. *British Journal of Mathematical Statistical Psychology*, 1974, *27*, 172–175.

Griffin, B. S., & Krutchkoff, R. G. Optimal linear estimators: An empirical Bayes version with application to the binomial distribution. *Biometrika*, 1971, *58*, 195–201.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

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

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

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Skibinsky, M. The maximum probability on an interval when the mean and variance are known. *Sankhya*, Series A, 1977, *39*, 144–159.

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.

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

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. *Educational and Psychological Measurement*, in press.

Acknowledgements

The author thanks two anonymous referees for their helpful comments. The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

Author's Address

Rand R. Wilcox, UCLA Graduate School of Education, Center for the Study of Evaluation, 145 Moore Hall, Los Angeles, CA 90024.