DOCUMENT RESUME

ED 138 642 TM 006 297

AUTHOR Wilcox, Rand
TITLE Estimating the Likelihood of False-Positive and False-Negative Decisions in Mastery Testing: An Empirical Bayes Approach.
PUB DATE [Apr 77]
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Computer Programs; *Error Patterns; Item Analysis; Mastery Learning; *Mastery Tests; Mathematical Models; Statistical Analysis; True Scores
IDENTIFIERS Monte Carlo Method

ABSTRACT

False-positive and false-negative dicisions are the fundamental errors committed with a mastery test; yet the estimation of the likelihood of committing these errors has not been investigated. Accordingly, two methods of estimating the likelihood of committing these errors are described and then investigated using Monte Carlo techniques. Conditions for obtaining accurate estimates are noted. (Author)

ESTIMATING THE LIKELIHOOD OF FALSE-POSITIVE AND FALSE-NEGATIVE DECISIONS IN MASTERY TESTING: AN EMPIRICAL BAYES APPROACH

Rand Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

A paper presented at the Annual Meeting of the American Educational Research Association. New York, April 1977.

ABSTRACT

False-positive and false negative decisions are the fundamental errors committed with a mastery test; yet the estimation of the likelihood of committing these errors has not been investigated. Accordingly, two methods of estimating the likelihood of committing these errors are described and then investigated using Monte Carlo techniques. Conditions for obtaining accurate estimates are noted.

ESTIMATING THE LIKELIHOOD OF FALSE-POSITIVE AND FALSE-NEGATIVE DECISIONS IN MASTERY TESTING: AN EMPIRICAL BAYES APPROACH

1. Introduction

Typically a mastery test is designed to sort k examinees into one of two mutually exclusive groups. For example, in a program of Individually Prescribed Instruction a student's progress through each level of a program of study is governed by his performance on a test dealing with individual behavioral objectives. The purpose of a test in such situations is to make a mastery/nonmastery decision for each of k examinees. If a mastery decision is made for a particular examinee then he is advanced to the next level of instruction. If, however, a nonmastery decision is made he will be given remedial work.

A model of mastery testing which is frequently adopted may be described as follows: A pool or domain of dichotomously scored test items, having mixed item difficulty, is constructed in relation to a particular course of instruction. The item pool may exist de facto or it may be a convenient conceptualization. The item form no[illegible] [illegible]ively and others (1973) represents such a conceptualization. Let λ_i [illegible]e the percent correct domain or "true" score of the i^{th} examinee; λ_i represents the percent of items that the ith examinee would answer correctly if he were to respond to every item in the item pool at a given occasion in time. With respect to the domain of items an examinee is said to have attained mastery if $\lambda_i \geq \lambda_o$ and nonmastery if $\lambda_i < \lambda_o$ where λ_o is a known constant with a value between zero and one. The problem is to make a mastery/nonmastery decision for a given examinee based on his responses to n items randomly selected from the item domain. A mastery decision is made if an examinee answers n_o or more items correctly where $0 \leq n_o \leq n$. Note that λ_o corresponds to a concept of mastery and n_o to a mastery score or index (Harris, 1974). We further observe that the model of mastery testing just described is equivalent to the ranking and selection problem of partitioning k populations (examinees) with respect to a standard.

This model provides a reasonable description of mastery testing and is consistent with definitions of mastery or criterion-referenced tests (Glaser and Nitko, 1971; Harris, 1974). (See also Hambleton and Novick, 1973; Fhaner, 1974; Novick and Lewis, 1974; Huynh, 1976; Wilcox, 1976).

A false-positive error occurs when the examiner estimates an examinee's true score λ_i to be above the criterion level λ_o when in fact it is not. A false-negative error occurs when λ_i is estimated to be below λ_o when the reverse is true. False-positive and false-negative errors are the two errors that can be made in a two-valued classification, yet the estimation of the probability of committing these types of errors in connection with mastery testing has been virtually ignored. Instead attention has been given to measures of stability such as the proportion of agreement (Hambleton and Novick, 1973) which estimates the probability of randomly selecting an examinee and classifying him the same way based on two administrations of the same test. Certainly it is desirable to have a test with a high degree of stability. However, it may be that such a test is consistently inaccurate.

Let α and β equal the probability of committing a false-positive and false-negative decision, respectively, for an examinee chosen at random from some population of potential examinees. Observe that the values of both α and β are a function of the number of items on the test as well as the instructional history of the examinees. Consequently, knowing α and β provides a meaningful characterization of the entire teaching-testing complex. The purpose of this paper is to examine the problem of estimating α and β based on student response data.

The binomial error model gives a reasonable approximation to the observed score distributions on tests (Lord, 1965, p. 253), but the compound binomial may be more realistic (Lord, 1965, Section 6; Lord and Novick, 1968, Chapter 23) and hence may give more accurate results. Accordingly, two methods of estimating α and β are described and the accuracy of these statistics are examined under both the binomial and compound binomial error model. Sections 3 and 4 derive estimates of α and β assuming that the distribution of true scores belongs to a particular parametric family. Section 5 examines the accuracy of these estimation procedures using Monte Carlo techniques.

2. Mathematical Statement of the Problem

As indicated earlier, we let λ_i denote the proportion correct "true" score of an examinee taking an n item test. We regard λ_i $(i=1,\ldots,k)$ as a sample from a prior distribution, say $g(\lambda)$, where $0 \leq \lambda \leq 1$. Let $h(x \mid \lambda_i)$ be the distribution of observed scores for a given true score λ_i. Assuming that $g(\lambda)$ is an integrable function and since $h(x \mid \lambda_i)$ is discrete we have:

$$(2.1) \qquad \alpha = \sum_{x=n_0}^{n} \int_0^{\lambda_0} h(x \mid \lambda)\, g(\lambda)\, d\lambda.$$

and

$$(2.2) \qquad \beta = \sum_{x=0}^{n_0-1} \int_{\lambda_0}^{1} h(x \mid \lambda)\, g(\lambda)\, d\lambda$$

If we knew $g(\lambda)$ and if we assume $h(x \mid \lambda)$ is binomial we would also know α and β. However, $g(\lambda)$ is usually unknown. The approach taken here is to use empirical Bayes procedures to estimate $g(\lambda)$. When $h(x \mid \lambda)$ is assumed to be compound binomial, it too must be estimated. We are particularly interested in the accuracy of point estimates of α and β for relatively small values of k and n. We emphasize that the results given below do not reflect directly the accuracy of our estimate of $g(\lambda)$. In fact we are only concerned with an accurate estimate of $g(\lambda)$ in so far as it improves our estimate of α and β. It may be, for example, that a relatively poor estimate of $g(\lambda)$ will yield a reasonably accurate estimate of the frequency of occurrence of both false-positive and false-negative decisions.

As a measure of the accuracy of any statistic $\hat{\alpha}$ which is used to estimate α we use the expected value of the square of the difference of α and $\hat{\alpha}$ over the joint distribution of x and λ. That is, we use

$$w_a^2 = \sum_{x=0}^{n} \int_0^1 (\alpha-\hat{\alpha})^2 \, h(x \mid \lambda) \, g(\lambda) \, d\lambda. \tag{2.3}$$

which corresponds to the average risk used in empirical Bayes methods (see, e.g., Maritz, 1970, p. 3). When estimating β with $\hat{\beta}$ we use w_b^2 which is defined by replacing $(\alpha-\hat{\alpha})^2$ with $(\beta-\hat{\beta})^2$ in (2.3).

3. Estimation of α and β assuming a beta prior

The estimation of the prior distribution $g(\lambda)$ is a most difficult problem for which no general solution exists. It behooves us, therefore, to consider the estimation of $g(\lambda)$ under a variety of conditions. We begin with perhaps the simplest, but also the most severe restriction on the family of prior distributions, namely, that $g(\lambda)$ is an incomplete beta distribution with parameters r and s. That is,

$$g(\lambda) = \frac{\Gamma(r + s)}{\Gamma(r)\,\Gamma(s)} \lambda^{r-1} (1-\lambda)^{s-1}$$

where Γ is the usual gamma function. This assumption is restrictive because there is little if any doubt that $g(\lambda)$ does not belong to the family of beta priors. Yet there are several reasons for considering this case. First, the beta density is the natural conjugate prior of the binomial kernal $h(x \mid \lambda)$ (Raiffa and Schlaifer, 1961). Second, we have indentifiability, i.e., there exists a unique g such that

$$f(x) = \int_0^1 h(x \mid \lambda)\, g(\lambda)\, d\lambda \tag{3.1}$$

where $f(x)$ is the marginal distribution of observed scores (Maritz, 1970, chapter 2). Third, there is evidence that a reasonable though not entirely satisfactory approximation of the true score distribution can be obtained with a two parameter beta prior (Keats and Lord, 1962). Finally, and perhaps most importantly of all, the results of estimation procedures assuming a particular parametric form for the prior may be used as a bench mark for judging alternate estimation techniques.

Let x_{ij} $(i=1,\ldots,k;\ j=1,\ldots,n)$ denote the j^{th} observation on the i^{th} examinee. By estimating the parameters r and s of the beta prior based on the sample x_{ij}, we obtain an estimate of g which in turn yields an estimate of both α and β. We begin by describing a method of estimating r and s which assumes that the binomial error model (Lord and Novick, 1968, chapter 23) holds, i.e.

$$h(x \mid \lambda) = \binom{n}{x} \lambda^{x} (1-\lambda)^{n-x}$$

Let $M_{[t]}$ denote that t^{th} factorial moment of the marginal distribution of observed scores, i.e.,

$$M_{[t]} = \sum_{x=0}^{n} \frac{x!}{(x - t)!} f(x)$$

Let μ_t represent the t^{th} moment of the true score distribution $g(\lambda)$. Then

(3.2) $$\mu_t = \frac{(n - t)!M_{[t]}}{n!}$$

(Lord and Novick, 1968, expression 23.8.4). We obtain unbiased estimates of $M_{[1]}$ and $M_{[2]}$ with

(3.3a) $$\hat{M}_{[1]} = 1/k \ \Sigma_{i=1}^{k} \ x_i$$

(3.3b) $$\hat{M}_{[2]} = 1/k \ \Sigma_{i=1}^{k} \ (x_i^2 - x_i)$$

where $x_i = \Sigma_{j=1}^{n} x_{ij}$. From (3.2) we have that

(3.4a) $$\hat{\mu} = \hat{M}_{[1]} /n$$

(3.4b) $$\hat{\mu}_2 = \hat{M}_{[2]} / (n(n - 1))$$

are unbiased estimates of μ_1 and μ_2. Thus, the mean, say μ and the variance, say σ^2, of the prior distribution may be estimated as

(3.5a) $$\hat{\mu} = \hat{\mu}_1$$

(3.5b) $$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$$

Note that it is possible to have $M_{[2]} = 0$ and $M_{[1]} > 0$ resulting in a negative estimate of σ^2. When this occurs we estimate each λ_i as

$$\hat{\lambda} = \Sigma_{i=1}^{k} \ \Sigma_{j=1}^{n} \ x_{ij}/(kn).$$

For this special condition, if $\hat{\lambda} \geq \lambda_o$ we estimate α to be zero and β to be $\sum_{x=0}^{n_o-1} \binom{n}{x}(\hat{\lambda})^x (1-\hat{\lambda})^{n-x}$. Correspondingly, if $\hat{\lambda} < \lambda_o$, we estimate β to be zero and α to be $\Sigma_{x=n_o}^{n} \binom{n}{x} (\hat{\lambda})^x (1-\hat{\lambda})^{n-x}$.

In terms of r and s

(3.6a) $$\mu = \frac{r}{r + s}$$

(3.6b) $$\sigma^2 = \frac{rs}{(r + s)^2 (r + s + 1)}$$

(Johnson and Kotz, 1970). Solving (3.6) for r and s gives

(3.7a) $$r = \frac{\mu^2 (1 - \mu)}{\sigma^2} - \mu$$

(3.7b) $$s = \frac{\mu (1 - \mu)}{\sigma^2} + \mu - 1$$

Substituting in (3.7) $\hat{\mu}$ and $\hat{\sigma}^2$ for μ and σ^2, respectively, yields an estimate of r and s, say $\hat{r}$ and $\hat{s}$. The estimates of r and s may then be used to estimate α and β as

(3.8a) $$\hat{\alpha}_1 = \sum_{x=n_o}^{n} \int_0^{\lambda_o} \binom{n}{x} \lambda^x (1-\lambda)^{n-x} \frac{\Gamma(\hat{r} + \hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

(3.8b) $$\hat{\beta}_1 = \sum_{x=0}^{n_o-1} \int_{\lambda_o}^{1} \binom{n}{x} \lambda^x (1-\lambda)^{n-x} \frac{\Gamma(\hat{r} + \hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

As indicated earlier, the binomial error model may not be completely satisfactory in an item sampling model. As suggested by Lord and Novick (1968, chapter 23) we use a two-term approximation to the compound binomial, viz., (3.9) $$\tilde{h}(x \mid \lambda) = p_n(x) + d\lambda(1-\lambda)\, C(x)$$

where

$$p_n(x) = \binom{n}{x} \lambda^x (1-\lambda)^{n-x}$$

$$C(x) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} p_{n-2}(x-v)$$

Lord (1965) notes that (3.9) is a close approximation to a frequency distribution for most cases of interest. Difficulties could arise if d were too large; we avoid these difficulties by assuming $0 \leq d \leq 4.0$. For all 16 distributions reported by Lord, the values of d were in this range. (See Lord, 1965, p. 264).

Under the more general compound binomial we are still able to estimate α and β. As shown by Lord (1965, p. 265) the mean and variance of the distribution of true scores for the two-term approximation to the compound binomial error model are given by:

(3.10) $$\mu = M_{[1]}/n$$

(3.11) $$\sigma^2 = \sigma_x^2 - (n-2d)\,\overline{p}\,\overline{q}$$

where σ_x^2 is the variance of the distribution of observed scores, $\overline{p} = M_{[1]}/n$ and $\overline{q} = 1-\overline{p}$. The parameter d is given by

(3.12) $$d = \frac{n^2 (n-1)\, \sigma_\pi^2}{2[\mu_x (n - \mu_x) - \sigma_x^2 - n\,\sigma_\pi^2]}$$

where σ_π^2 is the variance of the item difficulties. d may be estimated using standard item analysis techniques. Hence, the parameters r and s of the beta prior may be estimated using (3.7) above.

Substituting (3.9) into (3.8a), the estimate of α under the compound binomial error model is

$$(3.13a) \qquad \hat{\alpha}_{1(d)} = \sum_{x=n_o}^{n} \int_{0}^{\lambda_o} \tilde{h}(x \mid \lambda) \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

Correspondingly, we estimate β to be

$$(3.13b) \qquad \hat{\beta}_{1(d)} = \sum_{x=0}^{n_o-1} \int_{\lambda_o}^{1} \tilde{h}(x \mid \lambda) \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

4. Estimation of α and β using an inverse sine transformation.

In the previous section a procedure for estimating α and β was described which is contingent upon estimating the parameters r and s of an assumed beta prior. One difficulty with this estimation procedure is that the statistics $\hat{r}$ and $\hat{s}$ no doubt lack the desirable properties of unbiasedness, maximum likelihood, and efficiency. Consequently, one might expect estimates of r and s to be poor for relatively small samples. Since one would hope that accurate estimates of r and s would yield accurate estimates of α and β it may be helpful to search for more accurate estimates of r and s even though improvement in our estimates of r and s promises to be a most difficult task. For example, even in the simpler more conventional case in which the sampled values λ_i are known, maximum likelihood estimates of r and s are obtained iteratively. We propose, therefore, to investigate the use of an inverse sine transform which converts a binomial random variable into an approximately normally distributed random variable with <u>known</u> variance, the variance being independent of the value of λ_i. This is often called a variance stabilizing transformation. The advantage of this

approach is that the esti[illegible] parameters characterizing th[illegible] distribution can be expected to be more accurate relative to the beta-binomial model described above if the transformation used does indeed yield a normally distributed random variable. The disadvantage of this approach is that the transformed random variable is asymptotically normal and thus any estimation procedure using small samples may be poor. In addition, the rate of convergence to normality is a function of the unknown parameter λ_i. The crucial question is, of course, whether this approach reduces the values of w_a and w_b as defined by (2.3) above.

Let

$$(4.1) \qquad y_i = \tfrac{1}{2}(4n+2)^{\frac{1}{2}}\left(\sin^{-1}\left(\sqrt{\frac{x_i}{n+1}}\right) + \sin^{-1}\left(\sqrt{\frac{x_i+1}{n+1}}\right)\right)$$

where, as before, $x_i = \sum_{j=1}^{n} x_{ij}$ is the observed score of the i^{th} examinee. The transformation (4.1) is suggested by Freeman and Tukey (1950) where y_i given λ_i is approximately normally distributed with mean

$$\tau = (4n+2)^{\frac{1}{2}} \sin^{-1}(\sqrt{\lambda_i})$$

and variance one. As in the previous section we let μ and σ^2 represent the mean and variance of the prior distribution. Here, however, the natural conjugate prior has a normal distribution (Raiffa and Schlaifer, 1961). Moreover, the marginal distribution of observed score is also normally distributed with mean μ and variance $\sigma^2 + 1$. It follows that

$$E_\tau E(y \mid \tau) = E(\tau)$$

$$E_\tau E(y^2 - 1 \mid \tau) = E(\tau^2)$$

Hence, we may estimate $\mu = E(\tau)$ and $\sigma^2 = E(\tau^2) - E^2(\tau)$ with

$$\hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \tag{4.2a}$$

$$\hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{k} (y_i^2 - 1) - \hat{\mu}^2 \tag{4.2b}$$

It is known (see, for example, Hogg and Craig, 1970, p. 210) that the joint probability density function of y and λ is bivariate normal with common mean μ, respective variances $1 + \sigma^2$ and σ^2, and correlation $\sigma/\sqrt{1 + \sigma^2}$. Applying the method of moments, as was done in the previous section, the estimates of α and β are

$$\hat{\alpha}_2 = \int_{\gamma}^{\infty} \int_{-\infty}^{\lambda^*} \frac{1}{2\pi} \exp \left\{\frac{-(y-\lambda)^2}{2} - \frac{(\lambda-\hat{\mu})^2}{2\hat{\sigma}^2}\right\} d\lambda \, dy \tag{4.3a}$$

$$\hat{\beta}_2 = \int_{-\infty}^{\gamma} \int_{\lambda^*}^{\infty} \frac{1}{2\pi} \exp \left\{\frac{-(y-\lambda)^2}{2} - \frac{(\lambda-\hat{\mu})^2}{2\hat{\sigma}^2}\right\} d\lambda \, dy \tag{4.3b}$$

where

$$\gamma = \tfrac{1}{2}(4n+2)^{\frac{1}{2}} \left[\sin^{-1}\left(\sqrt{\frac{n_o}{n+1}}\right) + \sin^{-1}\left(\sqrt{\frac{n_o+1}{n+1}}\right)\right]$$

and

$$\lambda^* = (4n+2)\sin^{-1}(\sqrt{\lambda_o})$$

Again we have the difficulty that $\hat{\sigma}^2$ may be negative. In this case we set $\hat{\alpha}_2 = 0$, $\hat{\beta}_2 = \sum_{x=0}^{n_o-1} \binom{n}{x} \hat{\lambda}^x(1-\hat{\lambda})^{n-x}$ when $\hat{\lambda} \geq \lambda_o$; and when $\hat{\lambda} < \lambda_o$ we set $\hat{\beta}_2=0$, $\hat{\alpha}_2 = \sum_{x=n_o}^{n} \binom{n}{x} \hat{\lambda}^x (1-\hat{\lambda})^{n-x}$ as was done in the previous section.

It may be helpful to indicate how (4.3a) and (4.3b) can be evaluated with existing computer subroutines. For convenience we write the bivariate distribution of y and λ as $h(x \mid \lambda)\, g(\lambda)$. Observe that

$$(4.4) \qquad \mid \lambda)\, g(\lambda)\, d\lambda\, dy + \int_{-\infty}^{\gamma} \int_{\lambda'}^{\infty} h(x \mid \lambda)\, g(\lambda)\, d\lambda\, dy$$

$$= \int_{-\infty}^{\gamma} f(y)\, dy$$

where f(y) is the normally distributed, marginal distribution of the observed scores y with mean $\hat{\mu}$ and variance $1 + \hat{\sigma}^2$. The first integral on the left hand side of (4.4) can be evaluated with the IMSL subroutine (1975) MDBNOR after the random variables y and λ are transformed so as to have common mean zero and variance one. The right hand side of (4.4) can be evaluated with the FORTRAN subroutine ERFC; thus, we have the value $\hat{\beta}_2$. The statistic $\hat{\alpha}_2$ may be evaluated in a similar manner.

The transformation (4.1) is claimed by Mosteller and Youtz (1961) as well as Mosteller and Tukey (1968) to be the best existing angular transformation for the binomial distribution. As indicated above, however, the compound binomial may be a more appropriate probability distribution for describing the observed frequency of test scores. To be conservative we introduce an inverse sine transformation for the compound binomial. The desirability of using this transformation will be discussed in section 5 below.

As shown by Lord (1965, p. 265) the mean and variance of the two term approximation to the compound binomial distribution given by (3.12) may be written as $n\lambda$ and $(n-2d)\,\lambda\,(1-\lambda)$. It follows that

$$(4.5) \qquad \frac{2n}{\sqrt{n-2d}} \sin^{-1}(\sqrt{x/n})$$

is asymptotically normal with mean $\frac{2n \sin^{-1}(\sqrt{\lambda})}{\sqrt{n-2d}}$ and variance one (Rao, 1973, Section 6g). After estimating d via (3.11), one may use transformation (4.5) in place of (4.1), then estimate μ and σ^2 with (4.2a) and (4.2b), respectively, and ... e α and β with (4.3a) and (4.3b). When α and β are estimated using (4.5) for a given value of d, we denote the estimates by $\hat{\alpha}_{2(d)}$ and $\hat{\beta}_{2(d)}$, respectively.

5. Method and Results of Monte Carlo Experiments

The true score for each of the k examinees (the value of λ_i, i=1,...,k) was generated according to a beta distribution with parameters r and s. The priors used included J-shaped, U-shaped, symmetric and skewed distributions. Some of these distributions (e.g. U-shaped) are probably unrealistic in terms of mastery testing. They were included, however, so as to obtain more general results. Once λ_i was determined the observed score x_i was generated according to the two-term approximation to the compound binomial given by (3.9) for d=0.0, 2.5, 4.0. Expression (3.9) was evaluated by using the relationship $I_\lambda(a, n-a+1) = \sum_{x=a}^{n} \lambda^x (1-\lambda)^{n-x}$ in conjunction with IBM's SSP (1971) subroutine BDTR where I_λ denotes the incomplete beta function ratio (see Johnson and Kotz, 1970, chapter 24). Over 200 Monte Carlo studies were made for each of the estimators $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{\alpha}_2$ and $\hat{\beta}_2$.

Initially we set λ_o=0.7 and $n_o=\lambda_o n$. For each prior distribution used, w_a and w_b were estimated by first using the exact value of d and then by setting d arbitrarily equal to zero. The values of k and n were (k, n) = (10, 10), (10, 20), (10, 30), (20, 10), (20, 20), (30, 10). All estimates of w_a and w_b were based on 500 iterations. For simplicity we discuss the results in terms of w_a. No additional insights were found when examining w_b.

Regardless of the true score distribution used, the value of d had negligible effect on the value of w_a when (3.8a) was used to estimate α. This result is illustrated in Table I for the special case (r, s) = (9, 2) and (r, s) = (3, 3), λ_o = .7. Moreover, using the exact value of d in (3.13a) generally had little effect on lowering w_a as demonstrated in Table II. One exception to this finding occured for k=n=10 and r=s=3. For d=0, w_a was estimated to be .093 using (3.8a). For d=4.0, w_a dropped to .079. In general, however, varying the value of d affected w_a only at the third decimal place. This result also held when the more general (3.13a) was used which incorporates the two-term approximation to the compound binomial. This finding was not surprising since there appears to be negligible change in the observed score and true score distributions when the value of d is altered (Lord, 1965).

As [illegible] α_2, again it was found that altering d had little effect [illegible] the value of w_a. However, setting d arbitratily equal to zero and using transformation [illegible] tended to give better results (lower values of w_a) as opposed to using the exact value of d and transformation (4.5). The apparent reason for this [illegible] is that (4.1) converges more rapidly to normality than does (4.5).

Table I as well as Table II suggests that reasonably accurate estimates of α can be obtained particularly if n is greater than or equal to 30. To better assess the accuracy of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ we present Table III which gives the values of w_a for both $\hat{\alpha}_1$ and $\hat{\alpha}_2$ where r=9, s=2, λ_o=.5(.1).8. From Table III we see that $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are very accurate for λ_o=.5 but that this accuracy diminishes considerably as λ_o approaches .8. The difficulty is that both $\hat{\alpha}_1$ and $\hat{\alpha}_2$

tend to underestimate α. The results indicate that the amount by which α is underestimated increases as α gets large. In Table III, for example, the actual value of α is 0.005 when k=n=10 and λ_o=.5. For λ_o=.8, α=0.157.

Note that for λ_o=.8 the most effective method of lowering w_a is to increase n (the number of items) as opposed to increasing k (the number of individuals). In general, but not always, increasing n will decrease the value of α. Consequently, we lower the value of w_a by increasing n primarily because we obtain more accurate estimates of α when α is small. Increasing k with n fixed also lowered w_a but at a much slower rate.

We also observed that neither of the statistics $\hat{\alpha}_1$ or $\hat{\alpha}_2$ dominated the other, i.e., had consistently lower values for w_a. Consequently, based purely on statistical considerations, it is impossible to recommend one method of estimation rather than the other. However, we see that $\hat{\alpha}_1$ dominated $\hat{\alpha}_2$ for λ_o =.6, .7, .8 particularly for λ_o = .8. The reason is that $\hat{\alpha}_2$ = 0 occured more frequently due to negative estimates of the variance of the prior. Since the values of w_a were particularly large for λ_o = .8, it would seem best to use $\hat{\alpha}_1$. On the otherhand, to ensure accurate estimates of α, it would seem prudent to have n equal to at least 30 and preferably larger. In this case, evaluating $\hat{\alpha}_1$ might be more difficult computationally. By having n large, however, accurate estimates may still be possible with $\hat{\alpha}_2$.

TABLE I

Values of w_a using (3.8a), λ_0=.7

d\k,n	10, 10	10, 20	10, 30	20, 10	20, 20	30, 10
			r=9, s=2			
0.0	.043	.035	.031	.036	.028	.033
2.5	.04[illegible]	.035	.031	.036	.029	.034
4.0	.041	.035	.031	.036	.030	.034
			r=3, s=3			
0.0	.093	.059	.045	.087	.052	.087
2.5	.083	.057	.044	.081	.051	.081
4.0	.079	.056	.043	.077	.051	.077

TABLE II

Values of w_a using the exact value of d in (3.13a), λ_o=.7

d \ k,n	10, 10	10, 20	10, 30	20, 10	20, 20	30, 10
			r=9, s=2			
2.5	.041	.034	.029	.037	.029	.035
4.0	.040	.034	.029	.036	.030	.035
			r=3, s=3			
2.5	.083	.052	.038	.082	.051	.082
4.0	.081	.051	.038	.080	.050	.077

TABLE II

Values of w_a, $r=3$, $s=2$, $d=0$

λ_0 \ k,n	10, 10	10, 20	10, 30	20, 10	20, 20	30, 10
			Results using $\hat{\alpha}_1$			
.5	.017	.011	.008	.016	.011	.015
.6	.020	.014	.012	.018	.012	.016
.7	.043	.035	.031	.036	.028	.023
.8	.106	.082	.070	.094	.072	.091
			Results using $\hat{\alpha}_2$			
.5	.006	.004	.004	.004	.005	.005
.6	.019	.016	.014	.020	.016	.020
.7	.056	.035	.041	.058	.048	.058
.8	.127	.100	.085	.127	.100	.126

21

REFERENCES

Fhaner, Stig. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Freeman, M. F. and Tukey, J. W. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 1950, 21, 607-611.

Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement. (2nd ed.) Washington: American Council on Education, 1971.

Hambleton, R. K., and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

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

Hively, W., Maxwell, G., Rabehl, G., Sension, D. and Lundin, S. Domain-referenced curriculum evaluation: a technical handbook and a case study from the MINNEMAST project. Monograph No. 1, Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics. New York: Macmillan, 1970.

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

IBM Application Program, System/360. Scientific subroutines package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department, 1971.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass: Addison - Wesley, 1968.

Maritz, J. S. Empirical Bayes Methods London: Methuen, 1970.

Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronsen (Eds.) The Handbook of Social Psychology, Reading, Mass: Addison - Wesley, 1968.

Mosteller, F. and Youtz, C. Tables of the Freeman - Tukey transformations for the binomical and Poisson distributions. Biometrika, 1961, 48, 433-440.

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

Raiffa, H., and Schlaifer, R. Applied statistical decision theory. Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1961.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1973.

Wilcox, Rand R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.