
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. (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

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.
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 notably and others (1973) represents such a conceptualization. Let λ_i be the percent correct domain or "true" score of the i-th examinee; λ_i represents the percent of items that the i-th 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 λ_i ≥ λ_o and nonmastery if λ_i < λ_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 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 $\lambda_o$ 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; Paner, 1974; Novick and Lewis, 1974; Huynh, 1976; Wilcox, 1976).

A false-positive error occurs when the examiner estimates an examinee's true score $\lambda_i$ to be above the criterion level $\lambda_o$ when in fact it is not. A false-negative error occurs when $\lambda_i$ is estimated to be below $\lambda_o$ 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ are
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 \( \alpha \) and \( \beta \) 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 \( \alpha \) and \( \beta \) 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 \( \lambda_i \) denote the proportion correct "true" score of an examinee taking an \( n \) item test. We regard \( \lambda_i \) \( (i=1, \ldots, k) \) 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 \( \lambda_i \). Assuming that \( g(\lambda) \) is an integrable function and since \( h(x \mid \lambda_i) \) is discrete we have:

\[
(2.1) \quad \alpha = \sum_{x=n_0}^{n} \int_{0}^{\lambda_i} h(x \mid \lambda) g(\lambda) \, d\lambda.
\]

and

\[
(2.2) \quad \beta = \sum_{x=0}^{n_0-1} \int_{\lambda_i}^{1} h(x \mid \lambda) g(\lambda) \, d\lambda.
\]
If we knew \( g(\lambda) \) and if we assume \( h(x | \lambda) \) is binomial, we would also know \( \alpha \) and \( \beta \). However, \( g(\lambda) \) is usually unknown. The approach taken here is to use empirical Bayes procedures to estimate \( g(\lambda) \). When \( h(x | \lambda) \) is assumed to be compound binomial, it too must be estimated. We are particularly interested in the accuracy of point estimates of \( \alpha \) and \( \beta \) 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 \( \alpha \) and \( \beta \). 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 \( \alpha \) we use the expected value of the square of the difference of \( \alpha \) and \( \hat{\alpha} \) over the joint distribution of \( x \) and \( \lambda \). That is, we use

\[
(2.3) \quad w_a^2 = \sum_{x=0}^{n} \int_0^1 (\alpha - \hat{\alpha})^2 h(x | \lambda) g(\lambda) d\lambda.
\]

which corresponds to the average risk used in empirical Bayes methods (see, e.g., Maritz, 1970, p. 3). When estimating \( \beta \) 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 \( \alpha \) and \( \beta \) 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 \( \Gamma \) 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 kernel \( h(x | \lambda) \) (Raiffa and Schlaifer, 1961). Second, we have identifiability, i.e., there exists a unique \( g \) such that

\[
f(x) = \int_0^1 h(x | \lambda) \, g(\lambda) \, d\lambda
\]

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 benchmark 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 \( \alpha \) and \( \beta \). 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.

\[
h(x | \lambda) = \binom{n}{x} \lambda^x (1-\lambda)^{n-x}
\]
Let $M[t]$ denote the $t$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 $\mu_t$ represent the $t$th moment of the true score distribution $g(\lambda)$. Then

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

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

$$\hat{M}[1] = \frac{1}{k} \sum_{i=1}^{k} x_i$$

(3.3a)

$$\hat{M}[2] = \frac{1}{k} \sum_{i=1}^{k} (x_i^2 - \bar{x}_i)$$

(3.3b)

where $x_i = \frac{1}{n} x_{ij}$.

From (3.2) we have that

$$\hat{\mu}_1 = \frac{\hat{M}[1]}{n}$$

(3.4a)

$$\hat{\mu}_2 = \frac{\hat{M}[2]}{n(n-1)}$$

(3.4b)

are unbiased estimates of $\mu_1$ and $\mu_2$. Thus, the mean, say $\mu$ and the variance, say $\sigma^2$, of the prior distribution may be estimated as

$$\hat{\mu} = \hat{\mu}_1$$

(3.5a)

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$$

(3.5b)

Note that it is possible to have $M[2] = 0$ and $M[1] > 0$ resulting in a negative estimate of $\sigma^2$. When this occurs we estimate each $\lambda_i$ as

$$\hat{\lambda} = \frac{k}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} / (kn).$$
For this special condition, if $\lambda > \lambda_0$ we estimate $\alpha$ to be zero and $\beta$ to be

$$\sum_{x=0}^{n-1} \binom{n}{x} \lambda^x (1-\lambda)^{n-x}.$$  

Correspondingly, if $\lambda < \lambda_0$, we estimate $\beta$ to be zero and $\alpha$ to be

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

In terms of $r$ and $s$

(3.6a) \hspace{1cm} \mu = \frac{r}{r+s}

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

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

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

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

Substituting in (3.7) $\hat{\mu}$ and $\hat{\sigma}^2$ for $\mu$ and $\sigma^2$, 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 $\alpha$ and $\beta$ as

(3.8a) \hspace{1cm} \hat{\alpha}_1 = \sum_{x=n_0}^{n} \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r}) \Gamma(\hat{s})} \lambda^x (1-\lambda)^{n-x} d\lambda

(3.8b) \hspace{1cm} \hat{\beta}_1 = \sum_{x=0}^{n_0-1} \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r}) \Gamma(\hat{s})} \lambda^x (1-\lambda)^{n-x} 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.,
\( h(x \mid \lambda) = p_n(x) + d\lambda(1-\lambda) c(x) \)

where

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

\[
c(x) = 2 \sum_{v=0}^{\frac{2}{x}} (-1)^{v+1} \binom{2}{v} p_{n-2}(x-v)
\]

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 \( \alpha \) and \( \beta \). 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:

\[
\mu = M_{[1]} / n
\]

\[
\sigma^2 = \sigma_x^2 - (n-2d) \bar{p} \bar{q}
\]

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

\[
d = \frac{n^2 (n - 1) \sigma_n^2}{2[n \sigma_x(n - \mu_x) - \sigma_x^2 - \bar{n} \sigma_n^2]}
\]

where \( \sigma_n^2 \) 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 $\alpha$ under the compound binomial error model is

$$
\hat{\alpha}(d) = \sum_{x=0}^{n_0} \int_{\lambda_0}^{\lambda} h(x \mid \lambda) \frac{\Gamma(\hat{r} + \hat{s})}{\Gamma(\hat{r}) \Gamma(\hat{s})} \lambda^x (1-\lambda)^{\hat{s}-1} d\lambda
$$

Correspondingly, we estimate $\beta$ to be

$$
\hat{\beta}(d) = \sum_{x=0}^{n_0-1} \int_{\lambda_0}^{\lambda} h(x \mid \lambda) \frac{\Gamma(\hat{r} + \hat{s})}{\Gamma(\hat{r}) \Gamma(\hat{s})} \lambda^{x+1} (1-\lambda)^{\hat{s}-1} d\lambda
$$

4. Estimation of $\alpha$ and $\beta$ using an inverse sine transformation.

In the previous section a procedure for estimating $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$, 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 $\lambda_i$ are known, maximum likelihood estimates of $r$ and $s$ are obtained iteratively. We propose, therefore, to investigate the use of an inverse sine transformation which converts a binomial random variable into an approximately normally distributed random variable with known variance, the variance being independent of the value of $\lambda_i$. This is often called a variance-stabilizing transformation. The advantage of this
The approach is that the estimates of the parameters characterizing the 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 \( \lambda_i \). The crucial question is, of course, whether this approach reduces the values of \( \omega_a \) and \( \omega_b \) as defined by (2.3) above.

Let

\[
y_i = \lambda_i (4n + 2)^{\frac{1}{2}} \left( \sin^{-1} \left( \sqrt{\frac{x_i}{n + 1}} \right) + \sin^{-1} \left( \frac{x_i + 1}{n + 1} \right) \right)
\]

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 \( \lambda_i \) 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 \( \mu \) and \( \sigma^2 \) 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 \( \mu \) 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
\begin{align}
(4.2a) \quad \hat{\mu} &= \frac{1}{k} \sum_{i=1}^{k} \mathbf{y}_i \\
(4.2b) \quad \hat{\sigma}^2 &= \frac{1}{k} \sum_{i=1}^{k} (\mathbf{y}_i^2 - 1) - \hat{\mu}^2
\end{align}

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

\begin{align}
(4.3a) \quad \hat{\alpha}_2 &= \gamma \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp \left\{ -\frac{(\mathbf{y}\!-\!\hat{\mu})^2}{2}\right\} d\mathbf{y} \\
(4.3b) \quad \hat{\beta}_2 &= \gamma \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp \left\{ -\frac{(\mathbf{y}\!-\!\hat{\mu})^2}{2}\right\} d\mathbf{y}
\end{align}

where

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

and

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

Again we have the difficulty that \( \hat{\sigma}^2 \) may be negative. In this case we set

\[
\hat{\sigma}_2 = \begin{cases} 
0 & \text{if } \lambda \geq \lambda_0 \\
\sum_{x=0}^{n_o-1} (\mathbf{1}) \lambda^x (1-\hat{\lambda})^{n-x} & \text{when } \lambda < \lambda_0 \end{cases}
\]

\[
\hat{\beta}_2 = \sum_{x=n_o}^{\infty} (\mathbf{1}) \lambda^x (1-\hat{\lambda})^{n-x}
\]

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 $\lambda$ as $h(x \mid \lambda) g(\lambda)$. Observe that

$$
(4.4) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y) dy \
= \int f(y) dy
$$

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 $\lambda$ 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 $\hat{\sigma}_2$ 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

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

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 \( \nu \) and \( \sigma^2 \) with (4.2a) and (4.2b), respectively, and estimate \( \alpha \) and \( \beta \) with (4.3a) and (4.3b). When \( \alpha \) and \( \beta \) 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 \( \lambda_i, i=1,\ldots,k \)) was generated according to a beta distribution with parameters \( r \) and \( s \). The priors used included L-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 \( \lambda_i \) 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

\[
I_\lambda(a, n-a+1) = \sum_{x=a}^{\infty} \lambda^x (1-\lambda)^{n-x}
\]

in conjunction with IBM's SSP (1971) subroutine BDTR where \( I_\lambda \) 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 \( \lambda_0=0.7 \) and \( n_0=\lambda_0 n \). 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 \( \alpha \). This result is illustrated in Table I for the special case \((r, s) = (9, 2)\) and \((r, s) = (3, 5), \lambda_o = .7\). 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 occurred 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 it appears to be negligible change in the observed score and true score distributions when the value of d is altered (Lord, 1968).

As \( \lambda_o \), again it was found that altering d had little effect on the value of \( w_a \). However, setting d arbitrarily equal to zero and using transformation (4.5) 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 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 \( \alpha \) 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 \( w_{\hat{\alpha}_1} \) and \( w_{\hat{\alpha}_2} \) where \( r=9, s=2, \lambda_o = .5(1.8). \) From Table III we see that \( \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \) are very accurate for \( \lambda_o = .5 \) but that this accuracy diminishes considerably as \( \lambda_o \) approaches .8. The difficulty is that both \( \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \)
tend to underestimate $\alpha$. The results indicate that the amount by which $\alpha$ is underestimated increases as $\alpha$ gets large. In Table III, for example, the actual value of $\alpha$ is 0.005 when $k=n=10$ and $\lambda_0=.5$. For $\lambda_0=.8$, $\alpha=0.157$.

Note that for $\lambda_0=.8$ 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 $\alpha$. Consequently, we lower the value of $w_a$ by increasing $n$ primarily because we obtain more accurate estimates of $\alpha$ when $\alpha$ 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 $\lambda_0=.6, .7, .8$ particularly for $\lambda_0=.8$. The reason is that $\hat{\alpha}_2=0$ occurred more frequently due to negative estimates of the variance of the prior. Since the values of $w_a$ were particularly large for $\lambda_0=.8$, it would seem best to use $\hat{\alpha}_1$. On the otherhand, to ensure accurate estimates of $\alpha$, 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 $\hat{\alpha}_2$. 

18
### TABLE I

Values of \( w_a \) using (3.8a), \( \lambda_o = .7 \)

<table>
<thead>
<tr>
<th>( d ) ( k, n )</th>
<th>10, 10</th>
<th>10, 20</th>
<th>10, 30</th>
<th>20, 10</th>
<th>20, 20</th>
<th>30, 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 9, s = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.043</td>
<td>0.035</td>
<td>0.031</td>
<td>0.036</td>
<td>0.028</td>
<td>0.033</td>
</tr>
<tr>
<td>2.5</td>
<td>0.041</td>
<td>0.035</td>
<td>0.031</td>
<td>0.036</td>
<td>0.029</td>
<td>0.034</td>
</tr>
<tr>
<td>4.0</td>
<td>0.041</td>
<td>0.035</td>
<td>0.031</td>
<td>0.036</td>
<td>0.030</td>
<td>0.034</td>
</tr>
</tbody>
</table>

| \( r = 3, s = 3 \) |
| 0.0             | 0.093  | 0.059  | 0.045  | 0.087  | 0.052  | 0.087  |
| 2.5             | 0.083  | 0.057  | 0.044  | 0.081  | 0.051  | 0.081  |
| 4.0             | 0.079  | 0.056  | 0.043  | 0.077  | 0.051  | 0.077  |
TABLE II

Values of $w_a$ using the exact value of $d$ in (3.13a), $\lambda_d = 0.7$

<table>
<thead>
<tr>
<th></th>
<th>$k,n$</th>
<th>10, 10</th>
<th>10, 20</th>
<th>10, 30</th>
<th>20, 10</th>
<th>20, 20</th>
<th>30, 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r=9, s=2$</td>
<td>$d=2.5$</td>
<td>0.041</td>
<td>0.034</td>
<td>0.029</td>
<td>0.037</td>
<td>0.029</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>$d=4.0$</td>
<td>0.040</td>
<td>0.034</td>
<td>0.029</td>
<td>0.036</td>
<td>0.030</td>
<td>0.035</td>
</tr>
<tr>
<td>$r=3, s=3$</td>
<td>$d=2.5$</td>
<td>0.083</td>
<td>0.052</td>
<td>0.038</td>
<td>0.082</td>
<td>0.051</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>$d=4.0$</td>
<td>0.081</td>
<td>0.051</td>
<td>0.039</td>
<td>0.080</td>
<td>0.050</td>
<td>0.077</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>$k, n$</td>
<td>10, 10</td>
<td>10, 20</td>
<td>10, 30</td>
<td>20, 10</td>
<td>20, 20</td>
<td>30, 10</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results using $\hat{a}_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.017</td>
<td>.011</td>
<td>.008</td>
<td>.016</td>
<td>.011</td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td>.020</td>
<td>.014</td>
<td>.012</td>
<td>.018</td>
<td>.012</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td>.7</td>
<td>.033</td>
<td>.035</td>
<td>.031</td>
<td>.036</td>
<td>.028</td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td>.8</td>
<td>.106</td>
<td>.082</td>
<td>.070</td>
<td>.094</td>
<td>.072</td>
<td>.091</td>
<td></td>
</tr>
<tr>
<td>Results using $\hat{a}_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.006</td>
<td>.004</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td>.019</td>
<td>.016</td>
<td>.014</td>
<td>.020</td>
<td>.016</td>
<td>.020</td>
<td></td>
</tr>
<tr>
<td>.7</td>
<td>.056</td>
<td>.035</td>
<td>.041</td>
<td>.058</td>
<td>.048</td>
<td>.058</td>
<td></td>
</tr>
<tr>
<td>.8</td>
<td>.127</td>
<td>.100</td>
<td>.085</td>
<td>.127</td>
<td>.100</td>
<td>.126</td>
<td></td>
</tr>
</tbody>
</table>
REFERENCES

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


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


