Estimating the Parameters of Emrick's Mastery Testing Model

Wim J. van der Linden
Twente University of Technology

Emrick's model is a latent class or state model for mastery testing that entails a simple rule for separating masters from nonmasters with respect to a homogeneous domain of items. His method for estimating the model parameters has only restricted applicability inasmuch as it assumes a mixing parameter equal to .50 and an a priori known ratio of the two latent success probabilities. The maximum likelihood method is also available but yields an intractable system of estimation equations which can only be solved iteratively. The emphasis in this paper is on estimates to be computed by hand but nonetheless accurate enough for most practical situations. It is shown how the method of moments can be used to obtain such "quick and easy" estimates. In addition, an endpoint method is discussed that assumes that the parameters can be estimated from the tails of the sample distribution. A monte carlo experiment demonstrated that for a great variety of parameter values, test lengths, and sample sizes, the method of moments yields excellent results and is uniformly much better than the endpoint method.

In many instructional systems organized according to principles derived from recent developments in educational technology, the testing procedures being applied are criterion referenced. The major reason for applying these procedures, and not their norm-referenced counterparts, ordinarily lies in the fact that they utilize test items constructed on the basis of well-defined learning objectives. They thereby enable the test user to interpret test scores in terms of the specific knowledge and skills the student does and does not master. Norm-referenced measurements lack these properties; they are mainly of importance when the interest is in the relative standing of students in some norm group or population. Further differences between norm-referenced and criterion-referenced measurement are elucidated in Hambleton, Swaminathan, Algina, and Coulson (1978) and in van der Linden (in press).

When used for deciding whether the student has reached the learning objectives and may proceed with the next instructional unit (or take up a new course), criterion-referenced tests are ordinarily called mastery tests. Typically, a mastery decision is based on a cutoff score or mastery score on the test. Students with observed test scores exceeding this cutoff score are granted mastery status and discharged from the instructional unit. The others are the nonmasters; they are retained and in most cases they receive extra learning time or remedial instruction to enable them to reach the learning objectives.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 4, Fall 1981, pp. 517–530

0146-6216/81/040517-14$1.70

With regard to the student's true status underlying his or her observed test score, all existing mastery testing models can be classified as either *continuum* or *state models* (Meskauskas, 1976). Models of the former type postulate a latent or true score continuum, θ, underlying the observed test score and assume that a point θ_c, the true mastery score, is given dividing the continuum into a mastery ($\theta \geq \theta_c$) and a nonmastery ($\theta < \theta_c$) region. State models differ from continuum models in that they do not postulate a latent continuum but conceive mastery and nonmastery as two latent classes, each characterized by a different probability of a successful response to the test items. Ideally, this probability would be equal to one for a master and to zero for a nonmaster; but the influence of extraneous factors (measurement error) introduces a bias, making both probabilities differ from these ideal values. According to the state conception, mastery testing consists of collecting item responses generated by the two latent probabilities in an unknown ratio, estimating the values of these unknown parameters from the collected item responses, and classifying students as masters or nonmasters.

Emrick and Adams (1969) and Emrick (1971) were the first to introduce the latent class type of mastery testing model. Their terminology and approach is mainly Bayesian, and their model, generally referred to as Emrick's model, is in principle an application of the well-known binomial error model (Lord & Novick, 1968, chap. 23). Assuming threshold loss, they also derive an optimal cutoff score for separating masters from nonmasters. Besel (1973), Dayton and Macready (1976), and Macready and Dayton (1977) have presented state models that are essentially extensions of Emrick's model, obtained either by allowing the parameters to vary across the items or by imposing a hierarchical structure on the responses.

In this paper the concern is chiefly with procedures for estimating the parameters in Emrick's mastery testing model and with the properties of the cutoff score that can be computed from these estimates. The emphasis will be on "quick and easy" estimates, that is, on estimates in principle to be computed by hand or with the aid of a simple calculator. These estimates can be used not only in the event of no computer being available (e.g., classroom applications) but also as suitable starting values for the more complex iterative procedures involved in, for example, maximum likelihood estimation. Emrick and Adams (1969) and Emrick (1971) have suggested such an estimation procedure; but their procedure is based on an impractical assumption, requires the presence of prior information, and is therefore only of restricted meaning. This procedure will be reviewed in the next section and its restrictions will be elucidated. Comparatively simple estimates can be obtained by the method of moments and by an "endpoint" method (Muench, 1936). The purpose of this paper is to discuss the suitability of applying these estimation procedures to Emrick's model and to present the results of a monte carlo investigation carried out to compare their statistical properties.

Emrick's Mastery Testing Model

Let X be a random variable representing the number-correct test score obtained by a test with length n, probability α that a nonmaster will give a successful reply to an item, and probability β that a master will do likewise. Emrick's model simultaneously applies the binomial probability function with success parameters α and β as follows:

$$\begin{cases} p(x|\bar{M}) = \binom{n}{x}\alpha^x(1-\alpha)^{n-x} \\ p(x|M) = \binom{n}{x}\beta^x(1-\beta)^{n-x}, \end{cases} \qquad [1]$$

where M is a randomly drawn master and $\bar{M}$ a nonmaster. (In Emrick and Adams, 1969, and Emrick, 1971, $P(X \mid M)$ is given with $1 - \beta$ substituted for β; the parameterization in Equation 1 has been cho-

sen to be consistent with the more recent literature on state models for mastery testing). Although Emrick and Adams do not make this assumption explicitly, it is clear from their paper that

$$\alpha < \beta. \qquad [2]$$

For a population of students with Prob$\{M\} = \mu$, the model given in Equation 1 results in a distribution of test scores tending to be bimodal with modes at $n\alpha$ and $n\beta$.

The contribution of the above model lies not only in the possibility of using the (estimated) parameters α, β, and μ for theoretical and practical purposes, but also in an elegant decision rule that can be derived for granting mastery and nonmastery status to students. This rule is a monotone, nonrandomized Bayes rule, i.e., it has the form of a cutoff score c such that students with $X > c$ are declared to be masters, those with $X < c$ nonmasters, and the value of c is chosen to minimize the Bayes risk for a given population of students (for decision rules of this form, see, e.g., Ferguson, 1967, chap. 6, pp. 30-31). To arrive at this optimal rule, Emrick and Adams (1971) assume a threshold loss function with positive losses (which may be different) for the two incorrect decisions and zero loss for the two correct decisions:

$$L = \begin{cases} \ell_1, & \text{for a latent master with } X < c; \\ \ell_2, & \text{for a latent nonmaster with } X \geq c; \\ 0, & \text{otherwise.} \end{cases} \qquad [3]$$

The Bayes risk or expected loss for cutoff score c, which will be denoted by $B(c)$, is for the model of Equation 1 and the loss function of Equation 3 equal to

$$B(c) = \ell_1 \mu \sum_{x=0}^{c-1} \binom{n}{x} \beta^x (1-\beta)^{n-x} + \ell_2 (1-\mu) \sum_{x=c}^{n} \binom{n}{x} \alpha^x (1-\alpha)^{n-x} \qquad [4]$$

Completing the first sum,

$$B(c) = \ell_1 \mu \sum_{x=0}^{n} \binom{n}{x} \beta^x (1-\beta)^{n-x} - \sum_{x=c}^{n} \binom{n}{x} \left[\ell_1 \mu \beta^x (1-\beta)^{n-x} - \ell_2 (1-\mu) \alpha^x (1-\alpha)^{n-x} \right]. \qquad [5]$$

Since the binomial probability function has a monotone likelihood ratio with respect to x (Ferguson, 1967, sec. 5.2), it follows from Equation 2 that the bracketed factor in the second sum is negative up to some value of x, and positive thereafter. Therefore, Equation 5 is minimal if c assumes the value c^* satisfying

$$\ell_1 \mu \beta^{c^*} (1-\beta)^{n-c^*} - \ell_2 (1-\mu) \alpha^{c^*} (1-\alpha)^{n-c^*} = 0. \qquad [6]$$

This may be reduced to

$$c^* = \frac{\left[\ln \frac{1-\beta}{1-\alpha} + \frac{1}{n} \ln \frac{\lambda \mu}{1-\mu} \right] n}{\ln \frac{\alpha(1-\beta)}{(1-\alpha)\beta}}, \qquad [7]$$

where λ is equal to the loss ratio ℓ_1/ℓ_2. (For a somewhat different derivation, see Emrick and Adams, 1969.) In order to prevent Equation 7 from being indeterminate, the restrictions $\alpha > 0$, $\beta < 1$, $\alpha \neq \beta$ (this one is automatically satisfied by the inequality given in Equation 2), and $\mu \neq 1$ must be imposed. The last two restrictions also have some right in their own; they avoid degeneration of the model into a single binomial probability function.

Davis, Hickman, and Novick (1973, pp. 32-47) introduced and discussed the same model, deriving the same decision rule without realizing that these are Emrick's model and decision rule. Fricke (1974) discussed Emrick's model, providing tables for c^* as function of n, α, and β under the assumptions $\mu = .50$ and $\ell_1 = \ell_2$, and giving corrections to use in the event that one of these two assumptions is not met. Interestingly, the correction for $\mu \neq .50$ is independent of n and $\lambda = \ell_1/\ell_2$, whereas the correction for $\ell_1 \neq \ell_2$ does not depend on n and μ. The state model given by Besel (1973) is an extension of Emrick's model and allows for varying success probabilities across the items for both masters and nonmasters. Besel's model can be conceived as a simultaneous application of the compound binomial model (Lord & Novick, 1968, sec. 23.10; Walsh, 1953, 1959, 1963) with both latent classes characterized by a different set of parameter values.

Macready and Dayton (1977) present a response vector form of Emrick's model and use the maximum likelihood technique to obtain estimates for the parameters of the model. They also present a more general model, which is equivalent to Besel's model, and outline as well how maximum likelihood estimates for this model can be obtained. The possibility of incorporating a priori hierarchical relations between items into the two above state models and using them for validating behavioral hierarchies are given by Dayton and Macready (1976). References to other state models are Bergan, Cancelli, and Luiten (1980), Dayton and Macready (1980), Harris and Pearlman (1978), Knapp (1977), and Wilcox (1977a, 1977b, 1979a, 1979b). An excellent review of state models for mastery testing is given in Macready and Dayton (1980a).

Estimating α, β, and μ

Assuming $\mu = .50$ and an a priori known ratio of the two latent success probabilities, Emrick and Adams (1969) and Emrick (1971) have shown how estimators for α and β can be obtained via the square root of the interitem correlation. The fact that μ is assumed to take the value .50 (which is only mentioned in Emrick and Adams, 1969, and has also been documented by Wilcox and Harris, 1977) seriously restricts the applicability of this estimation method. Moreover, there will be hardly any situations in which the ratio α/β is a priori known.

As indicated earlier, one of the Macready and Dayton models is formally identical to Emrick's model. Their computer program MODEL3G can therefore be used to obtain maximum likelihood estimates of Emrick's parameters (Dayton & Macready, 1977; see also Macready & Dayton, 1980b). Though these estimates are attractive from a statistical point of view, maximum likelihood estimates for Emrick's model are not simple. The estimation equations do not yield closed-form estimators, and the iterative procedure used in MODEL3G (method of scoring) is too involved to be executed by hand. The same holds for Goodman's proportional fitting algorithm for obtaining maximum likelihood estimates (Goodman, 1974, 1975, 1979). This is also an iterative procedure, which generally has excellent properties and always converges (although, in some circumstances, convergence may be slow, Goodman, 1979, and not necessarily to maximum likelihood estimates, Goodman, 1974). For test lengths and sample sizes ordinarily encountered in mastery testing, the use of the proportional fitting algorithm requires access to a computer and is too involved to be used by hand. Properties of both the method of scoring and Goodman's algorithm for a constrained version of Emrick's model have been investigated in a monte carlo study by Houang and Harris (1980).

In this paper the emphasis is on "quick and easy" estimates, to be computed by hand or eventually with the aid of a simple calculator. Nevertheless, the estimators must be applicable in a wide variety of situations and have favorable statistical properties. An additional application of these estimates is their use as suitable start values in situations where facilities to employ the above iterative procedures for maximum likelihood estimation do exist. It is a common experience that the choice of good initial estimates reduces the required number of iterations considerably and prevents the procedure from converging to suboptimal "solutions." Before embarking on the introduction of simple estimators, note that in the following, closed-form estimators obtained under restrictions on test length have been left out of consideration. Examples are maximum likelihood estimators for the case of $n = 2$ (Wilcox, 1977a, 1977b) and the estimators derived in Werts, Linn, and Jöreskog (1973) for $n = 3$.

Method of Moments

Written as

$$p(x) = (1-\mu)\binom{n}{x}\alpha^x(1-\alpha)^{n-x} + \mu\binom{n}{x}\beta^x(1-\beta)^{n-x}, \quad [8]$$

with μ as an explicit parameter, it is clear that Emrick's model can be conceived as a *mixture* of two binomial distributions applied to the mastery testing problem. Mixtures of distributions (also called *compound* or *composite* distributions, but the term *mixture* has become current) have been extensively studied in the statistical literature. The properties of mixtures of two distributions have especially been subject of study, including well-known distributions as the normal (e.g., Rao, 1948; for a review, see Molenaar, 1965), the Poisson (Blischke, 1963; Rider, 1961), and the binomial (Blischke, 1962, 1963, 1964; Muench, 1936, 1938; Pearson, 1915; Rider, 1961). The last references have not been noted in the literature about state models for mastery testing, nor the fact that the method of moments has been used on several independent occasions to obtain estimators for mixtures of two binomials with results applying to Emrick's model.

For distributions with r unknown parameters, the method of moments consists of equating r moments (preferably the first r) to their corresponding sample functions and solving this system of equations for the parameters, provided, of course, that these r moments exist. Moment estimators are usually comparatively simple and are consistent under very mild conditions (Rao, 1973, p. 351).

Assuming Equation 2 and denoting the sample size by m, it can be shown that for a mixture of two binomials, the moment estimators are equal to

$$\hat{\alpha} = \frac{1}{2}A - \frac{1}{2}(A^2 - 4AF_1 + 4F_2)^{1/2}; \quad [9]$$

$$\hat{\beta} = \frac{1}{2}A + \frac{1}{2}(A^2 - 4AF_1 + 4F_2)^{1/2}; \quad [10]$$

$$\hat{\mu} = \frac{F_1 - \hat{\alpha}}{\hat{\beta} - \hat{\alpha}}, \quad [11]$$

with

$$F_k = \frac{1}{m}\sum_{j=k}^{n} \frac{j(j-1)\ldots(j-k+1)}{n(n-1)\ldots(n-k+1)} U_j; \quad [12]$$

$$A = \frac{F_3 - F_1F_2}{F_2 - F_1^2}, \qquad [13]$$

and the restriction

$$0 < (A^2 - 4AF_1 + 4F_2)^{1/2} \leq \min\{A, 2-A\} \qquad [14]$$

(Blischke, 1962, 1964; see also Johnson & Kotz, 1969, sec. 3.11). Equation 12 is, up to the factor $(n-1) \ldots (n-k+1)$, equal to the definition of the k^{th} sample factorial moment, where U_j denotes the number of students in the sample with test score $x = j$ ($j = 0, \ldots, n$). The restriction in Equation 14 is necessary to obtain solutions that are real valued and in the interval [0, 1].

Although moment estimators usually have an asymptotic relative efficiency less than 1, Blischke (1962) has shown that a mixture of two binomials is an exception and that Equations 9 through 11 do have an asymptotic efficiency equal to 1. This means that for $n \to \infty$ the efficiency of these estimators tends to be equal to that of maximum likelihood estimators, which is the Cramér-Rao lower bound. He has also shown that the limiting joint distribution of Equations 9 through 11 is the normal, with first marginal moments equal to α, β, and μ, respectively, and a known, albeit tedious to compute, variance-covariance matrix.

For data sets generally encountered in mastery testing, Equation 12 can be computed easily by hand for $k = 1, 2, 3$. Once A has been calculated from Equation 13, Equations 9 through 11 give the desired estimates.

The "Endpoint" Method

Reulecke (1977a) discusses Emrick's model and proposes a method of estimation in which α is treated as an a priori known parameter and β and μ are estimated from the observed frequencies in the right-hand tail of the sample distribution. His method for estimating β is exactly the same as the one used by Muench (1936) for fitting mixtures of binomials to biological data; the terminology in this paper will therefore be derived from Muench and this method will be called the "endpoint" method.

It is assumed in Reulecke's method that the applicability of Emrick's model is restricted to multiple-choice items and that α can be treated as an a priori known parameter equal to q^{-1} (q being the number of alternatives).

To estimate β, the endpoint method assumes that the right-hand tail of the score distribution is virtually unmixed and that U_{n-1} and U_n can therefore be considered to come from a single binomial distribution with parameters β and n. Equating the ratio of these observed frequencies to the ratio of their estimated expected frequencies yields

$$\frac{U_n}{U_{n-1}} = \frac{\binom{n}{n}\tilde{\beta}^n(1-\tilde{\beta})^{n-n}}{\binom{n}{n-1}\tilde{\beta}^{n-1}(1-\tilde{\beta})}, \qquad [15]$$

which, on solving for the endpoint estimator $\tilde{\beta}$, results in

$$\tilde{\beta} = \frac{nU_n}{U_{n-1} + nU_n} \qquad [16]$$

(see also Reulecke, 1977b).

An estimate of μ is obtained by Reulecke via

$$\frac{U_n}{m\tilde{\mu}} = \binom{n}{n}\tilde{\beta}^n(1 - \tilde{\beta})^{n-n} \qquad [17]$$

as

$$\tilde{\mu} = U_n\left[\frac{U_{n-1} + nU_n}{nU_n}\right]^{n}_{m}{}^{-1}. \qquad [18]$$

Note that there is no reason to restrict the applicability of the endpoint method to the right-hand tail of the score distribution. When the items are not of the multiple-choice type, or the knowledge or random guessing model on which $\alpha = q^{-1}$ is based does not hold for other reasons, α can be estimated from the ratio of U_0 to U_1 as

$$\tilde{\alpha} = \frac{U_1}{nU_0 + U_1}. \qquad [19]$$

It is also possible, analogous to Equation 18, to derive a second endpoint estimator for μ from $\tilde{\alpha}$ and U_0. However, the problem this creates can be circumvented by using neither of the two and instead by substituting α and β in the first moment equation:

$$\dot{\mu} = \frac{F_1 - \tilde{\alpha}}{\tilde{\beta} - \tilde{\alpha}}. \qquad [20]$$

To some extent this also circumvents another problem of endpoint estimation, namely, that it uses only the observations in the two outermost categories and throws away the larger part of the data. Endpoint estimators are thus inefficient and likely to yield, for the same sample size, larger errors of estimation than estimators more fully exploiting the information in the sample. There is even a nonzero probability of $\tilde{\alpha}$ ($U_0 + U_1 = 0$), $\tilde{\beta}$ ($U_{n-1} + U_n = 0$), or $\tilde{\mu}$ ($U_n = 0$) being undefined. Especially for smaller samples and longer tests this might give rise to situations in which the endpoint method is unusable and, for example, moment or maximum likelihood estimation is still possible.

From Equations 16 and 19 it is clear that for all samples with $(U_0 + U_1) > 0$ or $(U_{n-1} + U_n) > 0$, $\tilde{\alpha}$ and $\tilde{\beta}$ are in the interval [0, 1]. A comparable property, however, does not hold for Equation 18. For example, when $n = 10$, $m = 100$, $U_9 = 20$, and $U_{10} = 6$ (these two frequencies have a joint probability of occurrence that increases as β and μ approach 1), it appears that $\tilde{\mu} = 1.07$. In such a case, it seems natural to equate $\tilde{\mu}$ to the maximum value of μ, that is put $\tilde{\mu} = 1$.

A Monte Carlo Experiment

The properties of the moment and endpoint estimators were examined and compared with each other, using monte carlo techniques. The two success parameters, the mixing parameter, test length, and sample size (number of students) were varied, and for each combination the expected error of estimation and the risk function using squared error loss were computed. This was done for both the model parameters to be estimated—α, β, and μ—and the optimal cutoff score, c^*. All results reported here are each based on 1,000 replications. Data were simulated with the aid of random procedures from the NAG Fortran Library (1977).

Table 1 shows the results for the moment estimators with varying α and β. The parameter values

Table 1
Results for Moment Estimators with Varying Success Parameters and $\mu = .70$, $n = 10$, and $m = 100$

(α, β)	(.10, .90)		(.25, .90)		(.25, .75)		(.25, .60)		(.40, .60)	
	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$
$\hat{\alpha}$	.000	.000	-.001	.001	-.002	.002	.004	.004	-.012	.016
$\hat{\beta}$	.000	.000	.000	.000	-.001	.001	-.003	.001	.021	.007
$\hat{\mu}$	.001	.002	.000	.002	.001	.004	.010	.012	-.149	.133
$\hat{c}^*_{.25}$	-.013	.067	-.001	.074	.029	.182	.001	.718	.104	10.468
$\hat{c}^*_1$	-.013	.064	.000	.071	-.029	.161	-.009	.548	.433	10.734
$\hat{c}^*_4$	-.012	.061	.000	.070	-.028	.145	-.025	.426	.760	11.843

single binomial is systematically varied. As will be seen below, this is a crucial factor in interpreting the properties of the various estimators. In all tables, ε is a generic symbol for an error of estimation. The results for $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\mu}$ are extremely good, both according to $E\varepsilon$ and $E\varepsilon^2$. (The latter is more informative, of course, since it contains the square of the former as an additive component). An exception must be made for $\hat{\mu}$ with $(\alpha, \beta) = (.40, .60)$. This estimator is clearly more dependent on the difference between the values for α and β than the estimators for these parameters themselves. Its results are even worse than may seem at first sight. During the experiment account was kept of the number of cases in which the inequalities in Equation 14 were not met. For the 12,000 replications in Tables 1 through 4, this happened only occasionally (131 times)[1].

Most of these cases occurred for $(\alpha, \beta) = (.40, .60)$, that is, for the situation in which Emrick's model approaches a single binomial. Whenever these inequalities were not met, Blischke's advice was followed and a single binomial was fitted with parameters $\hat{\alpha} = \hat{\beta} = F_1$ and $\hat{\mu} = 1$. With true parameter value μ, this can be expected to lead to a positive bias for $\hat{\mu}$ rather than the negative one shown in Table 1. ($E\varepsilon$ and $E\varepsilon^2$ could also have been computed only for those cases satisfying Equation 14, possibly eliminating such biases, but the line of action chosen here was expected to be closer to the one that will be pursued in practice.)

More important, perhaps, than the results for $\hat{\alpha}, \hat{\beta}$, and $\hat{\mu}$ are the results for $\hat{c}^*$. Table 1 gives these for a loss ratio $\lambda = \ell_1/\ell_2$ equal to .25, 1, and 4. It is important to note that these results should be evaluated against the test length $n = 10$ used in Table 1. Taking this into account, all results are excellent with the exception of the risk for (.40, .60). A risk from 10 to 12 amounts to a standard error of estimation between 3 and 4, and for a 10-item test this is too large to be practicable.

Results for the moment estimators with varying mixing parameters are given in Table 2. Going from .50 to .90, there is a slight loss in efficiency. On the whole, however, the results for both the parameter estimates and $\hat{c}^*$ are extremely good. The difference in bias and risk between $\hat{\alpha}$ and $\hat{\beta}$ for $\mu = .90$ may be explained by the fact that 12 of the 1,000 replications did not meet the inequalities in Equation 14, and that putting $\hat{\alpha} = \hat{\beta} = F_1$ for a data set with $\mu = .90$ introduces a larger bias in $\hat{\alpha}$ than in $\hat{\beta}$. No such replications were encountered for $\mu = .50$ and $\mu = .70$.

Table 3 shows how the results for the moment estimators vary with test length. The general impression is that the results are better the longer the test, and that even for a test as short as five items

[1]The center column is the same in Table 1 through 3 and serves as a benchmark.

Table 2
Results for Moment Estimators with Varying Mixing Parameter and $\alpha = .25$, $\beta = .75$, $n = 10$, and $m = 100$

μ	.50		.70		.90	
	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$
$\hat{\alpha}$	.001	.001	-.002	.002	.009	.011
$\hat{\beta}$	-.001	.001	-.001	.001	.002	.000
$\hat{\mu}$	.002	.004	.001	.004	-.020	.012
$\hat{c}^*_{.25}$	.003	.115	-.029	.182	.006	.570
$\hat{c}^*_1$	.000	.113	-.029	.161	-.004	.411
$\hat{c}^*_4$	-.004	.116	-.028	.145	-.015	.291

moment estimators are still usable for estimating Emrick's model and for computing its optimal cutoff score.

A comparable impression can be derived from Table 4, where results clearly show the previously mentioned asymptotic efficiency of the moment estimators. It appears that $\hat{c}^*$ shares this property for the three loss ratio values.

The endpoint estimators showed less optimistic results. Both the estimators given in Equation 16 and 18 through 20 and the optimal cutoff scores that can be computed from these were checked. From the great variety of parameter sets that were used, three were closer and are shown in Table 5 to illustrate how wildly the endpoint estimators fluctuate and how inefficient they are. Parameter Set I

Table 3
Results for Moment Estimators with Varying Test Length and $\alpha = .25$, $\beta = .75$, $\mu = .70$, and $m = 100$

n	5		10		20	
	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$
$\hat{\alpha}$	.001	.007	-.002	.002	.001	.000
$\hat{\beta}$	.003	.002	-.001	.001	.000	.000
$\hat{\mu}$	-.013	.012	.001	.004	.002	.002
$\hat{c}^*_{.25}$	.004	.301	-.029	.182	.009	.156
$\hat{c}^*_1$	.015	.246	-.029	.161	.007	.144
$\hat{c}^*_4$	.026	.209	-.028	.145	.005	.134

Table 4
Results for Moment Estimators with Varying Sample Size and $\alpha = .25$, $\beta = .75$, $\mu = .90$, and $n = 10$

m	25		50		500	
	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$
$\tilde{\alpha}$	.007	.010	-.002	.003	.000	.000
$\hat{\beta}$	.001	.002	.000	.001	.000	.000
$\hat{\mu}$	.021	.021	-.002	.007	-.001	.001
$\hat{c}^*_{.25}$	-.002	.859	-.031	.353	-.002	.032
$\hat{c}^*_1$	-.002	.738	-.029	.303	-.002	.028
$\hat{c}^*_4$	-.004	.650	-.028	.266	-.002	.024

($\alpha = .10, \beta = .90, \mu = .70, n = 10, m = 100$) is the set with the best results. Its results can be compared with those of the same parameter set in Table 1, but the moment estimators are nevertheless superior. Especially the optimal cutoff score, $\tilde{c}^*$, computed from $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\mu}$, for which the results are given in rows 5 through 7 of Table 5, is inferior to the cutoff score based on moment estimators, $\hat{c}^*$. It is interesting to note that the estimator $\dot{\mu}$ given in Equation 20, which is based on $\tilde{\alpha}$, $\tilde{\beta}$, and the first moment equation, displays a considerable improvement on $\tilde{\mu}$. As can be seen in the last three rows of Table 5, substituting $\dot{\mu}$ instead of $\tilde{\mu}$ into the optimal cutoff score results in a remarkable gain in efficiency. Parameter Set II ($\alpha = .25$, $\beta = .75$, $\mu = .70$, $n = 10$, $m = 25$) was one of the poorest sets. It shows risk values that cannot be tolerated in practice.

Results typical of what was normally encountered were obtained for Parameter Set III ($\alpha = .25$, $\beta = .75$, $\mu = .50$, $n = 10$, $m = 100$). The risk values are smaller than those for Set II but still too large for practical purposes. It is also seen that $\dot{\mu}$ is here again an improvement on $\tilde{\mu}$.

The above impression of the endpoint estimators is still too optimistic. Just as in the case of the moment estimators, account was kept of how often inadmissible endpoint estimates were met for the various parameter sets. Unlike the moment estimators, where only an overall percentage of inadmissible estimates hardly exceeding 1% was found, the endpoint estimators $\tilde{\mu}$ and $\dot{\mu}$ showed large numbers of inadmissible values. Test length and the value of μ proved to be especially critical. Percentages exceeding 20% or 30% were no exception. There was even a case with $\mu = .90$ in which some 47% of the replications yielded values for $\tilde{\mu}$ larger than one. (In all these cases the estimators were set equal to zero or one before entering the computations, so that Table 5 is based on these cases as well.)

A different problem was met with the endpoint parameters $\tilde{\alpha}$ and $\tilde{\beta}$. As can be seen from Equations 16 and 19, these expressions are indeterminate whenever the tails of the sample distributions are empty, that is, when $U_0 = U_1 = 0$ and/or $U_{n-1} = U_n$. This happened especially for α and β values close to each other, μ values close to one, small samples, and long tests. To give some examples: 8.8% for $\mu = .90$ with $\alpha = .25$, $\beta = .75$, $n = 10$, and $m = 100$; 15.9% for $m = 25$ with $\alpha = .25$, $\beta = .75$, and $n = 10$; and no fewer than 48% for $n = 20$ with $\alpha = .25, \beta = .75, \mu = .70$, and $m = 100$ (all percentages

Table 5
Some Results for the Endpoint Estimators

Parameter Set	I		II		III	
	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$	$E\varepsilon$	$E\varepsilon^2$
$\tilde{\alpha}$	.009	.002	.386	.346	.072	.053
$\tilde{\beta}$	-.004	.001	-.240	.219	-.062	.041
$\tilde{\mu}$	.025	.015	-.180	.140	.102	.124
$\dot{\mu}$	.001	.003	-.039	.119	.015	.085
$\tilde{c}^*_{.25}$	-.159	.842	-.920	6.512	-1.507	12.257
$\tilde{c}^*_1$	-.166	.853	-.617	5.614	-1.304	9.923
$\tilde{c}^*_4$	-.172	.868	-.287	4.726	-1.096	7.993
$\dot{c}^*_{.25}$	.017	.270	-.395	1.828	-.032	6.585
$\dot{c}^*_1$	.011	.254	-.168	1.633	-.009	6.352
$\dot{c}^*_4$	.004	.241	.058	1.582	.026	6.401

representing the number of times $U_0 = U_1 = 0$ was met during 1,000 replications). These cases have been left out of consideration in Table 5.

Conclusion

From the monte carlo experiments it can be concluded that the moment estimators do very well in nearly all situations. Best results were obtained for α and β values close to zero and one, respectively, μ values close to .50, long tests, and large samples. But even for 5-item tests and for sample sizes as small as 25 examinees, results were obtained that are sufficiently accurate for use in practice. This is a worthwhile property of moment estimators, since mastery tests are often said to be short and the sample of 25 examinees can be compared with a normal classroom size. Only one exception has to be made, namely, when the α and β values approach each other and when Emrick's model comes close to the model of a single binomial. In that case $\hat{\mu}$, and hence $\hat{c}^*$, lose their good properties and larger errors of estimation are likely to occur. The endpoint estimators proved to be unsuited for practical applications. They fluctuate too widely, yield unreliable cutoff scores, and were in all cases much inferior to the moment estimators.

Houang and Harris (1980) have reported on a monte carlo study of a constrained version of Emrick's model in which α is set equal to 0 (leaving only two parameters to be estimated). They compared the properties of four estimation procedures; two of these are Goodman's proportional fitting algorithm and the method of scoring for maximum likelihood estimation. Although the results were obtained for different sets of parameters values and are reported using different statistics, the impression is that for comparable sets of parameter values, these estimators behave comparably to the

moment estimators in this paper. In a companion paper (van der Linden, 1981b), moment estimators are derived for the case where α may be treated as a known parameter. A replication of the present experiment indicates that in this case the method of moment yields estimators that have most favorable properties and hardly get "upset" when the model comes close to a single binomial.

Elsewhere, the author of the paper has compared latent class and trait models for mastery testing and indicated that the latent class conception is akin to an all-or-none view of learning (van der Linden, 1978). He also proposed to reparameterize the latent class model and expressed some criticism of the interpretation of α and β as probabilities of guessing and forgetting, as given in Macready and Dayton (1977). However, this does not apply to the interpretation of α and β as just two different probabilities of a correct item response.

It is important to realize this when choosing a model for analyzing mastery tests. Emrick's model seems more realistic the more the instructional process results in a situation in which a mixture of two homogeneous groups of examinees with different success probabilities can indeed be expected. This situation will seldom be entirely reached. However, it may be that the educational tester is willing to tolerate this in exchange for the simplicity of Emrick's model, its elegant decision rule, and the fact that its parameters can be easily estimated. The choice of a psychometric model often depends on more factors than only the extent to which its assumptions are met.

During the monte carlo experiments it was observed that depending on test length, the optimal cutoff score pertaining to Emrick's model, c^*, has the desirable property of being robust with respect to the loss ratio. To illustrate this, a 20-item test with $\alpha = .25$, $\beta = .65$, and $\mu = .70$ yielded $c^* = 9.3$ for $\lambda = 2$, differing only slightly from the value $c^* = 9.9$ obtained when the loss ratio was reduced by a factor of 4 to $\lambda = \frac{1}{2}$. In practice, with test scores taking only integer values, this implies that for both loss ratios the same decisions will be made. This robustness follows from the fact that in Equation 7 the influence of λ is mitigated by the factor $1/n$. (For a discussion of the importance of robust decision rules with respect to loss functions and other models having this property, refer to van der Linden, 1980.)

In concluding, it is noted that other estimators for mixtures of binomial distributions, such as minimum χ^2 estimators and estimators by LeCam's procedure and Neyman's linearization technique, are available and have been examined by Blischke (1964). Blischke has also reported empirical results from a modest experiment in which, for only two sets of parameter values, the properties of minimum χ^2, LeCam's, and moment estimators for mixtures of two binomials were compared. That the results show considerably more bias and less efficiency for each of these estimators than has been found for the moment estimators in this paper may be due to the use of too small a number of iterations (namely, 105). The estimators discussed by Blischke have been left out of consideration here because they involve iterative computations and by no means yield "quick and easy" estimates. It is recalled, however, that the rate of convergence of such estimation procedures heavily depends on the quality of the initial guess used. The results in this paper demonstrate that moment estimates can excellently be used for this purpose as well.

References

Bergan, J. R., Cancelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. *Journal of Educational Statistics,* 1980, *5,* 65-82.

Besel, R. *Using group performance to interpret individual responses to criterion-referenced tests.* Paper presented at the annual meeting of the Ameri-

can Educational Research Association, New Orleans, April 1973. (ERIC Document Reproduction Service No. ED 076 650.)

Blischke, W. R. Moment estimators for the parameters of a mixture of two binomial distributions. *Annals of Mathematical Statistics,* 1962, *33,* 444-454.

Blischke, W. R. Mixtures of discrete distributions. In *Proceedings of the International Symposium on Discrete Distributions.* Montreal, 1963, 351-372.

Blischke, W. R. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association,* 1964, *59,* 510-528.

Davis, C. E., Hickman, J., & Novick, M. R. *A primer on decision analysis for individually prescribed instruction* (ACT Technical Bulletin No. 17). Iowa City IA: The American College Testing Program, 1973.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. *Psychometrika,* 1976, *41,* 189-204.

Dayton, C. M., & Macready, G. B. Model3G and Model5: Programs for the analysis of dichotomous, hierarchic structures. *Applied Psychological Measurement,* 1977, *1,* 412.

Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika,* 1980, *45,* 343-356.

Emrick, J. A. An evaluation model for mastery testing. *Journal of Educational Measurement,* 1971, *8,* 321-326.

Emrick, J. A., & Adams, E. N. *An evaluation model for individualized instruction* (Report RC 2674). Yorktown Heights NY: IBM, Thomas J. Watson Research Center, October 1969.

Ferguson, T. S. *Mathematical statistics: A decision theoretic approach.* New York: Academic Press, 1967.

Fricke, R. Zum Problem von Cut-off Formeln bei lehrzielorientierten Tests. *Unterrichtswissenschaften,* 1974, *3,* 43-56.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika,* 1974, *61,* 215-231.

Goodman, L. A. A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association,* 1975, *70,* 755-768.

Goodman, L. A. On the estimation of parameters in latent structure analysis. *Psychometrika,* 1979, *44,* 123-128.

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

Harris, C. W., & Pearlman, A. P. An index for a domain of completion or short answer items. *Journal of Educational Statistics,* 1978, *3,* 285-303.

Houang, R. T., & Harris, C. W. *Sampling variance of parameter estimates for a domain referenced latent class model.* Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Johnson, N. L., & Kotz, S. *Distributions in statistics: Discrete distributions.* Boston: Houghton Mifflin Company, 1969.

Knapp, T. R. The reliability of a dichotomous test item: A 'correlationless' approach. *Journal of Educational Measurement,* 1977, *14,* 237-252.

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

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics,* 1977, *2,* 99-120.

Macready, G. B., & Dayton, C. M. The nature and use of state mastery models. *Applied Psychological Measurement,* 1980, *4,* 493-516. (a)

Macready, G. B., & Dayton, C. M. A two-stage conditional estimation procedure for unrestricted latent class models. *Journal of Educational Statistics,* 1980, *5,* 129-156. (b)

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research,* 1976, *46,* 133-158.

Molenaar, W. Overzicht van ontmengingsmethoden voor twee normale verdelingen. *Statistica Neerlandica,* 1965, *19,* 249-263.

Muench, H. Probability distributions of protection test results. *Journal of the American Statistical Association,* 1936, *31,* 677-689.

Muench, H. Discrete frequency distributions arising from mixtures of several single probability values. *Journal of the American Statistical Association,* 1938, *33,* 390-398.

NAG Fortran Library. Manual—Mark 6. Oxford, England: Numerical Algorithms Group, 1977.

Pearson, K. On certain types of compound frequency distributions in which the components can be individually described by binomial series. *Biometrika,* 1915, *11,* 139-144.

Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society,* Series B, 1948, *10,* 159-203.

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

Reulecke, W. A. Ein Modell für die kriteriumorientierte Testauswertung. *Zeitschrift für Empirische Pädagogik,* 1977, *1,* 49-72. (a)

Reulecke, W. A. A statistical analysis of deterministic theories. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking and learning.* Bern, Switzerland: Huber, 1977. (b)

Rider, R. R. Estimating the parameters of mixed Poisson, binomial and Weibull distributions by the method of moments. *Bulletin de l'Institute Internationale Statistique,* 1961, *38,* 1-8.

van der Linden, W. J. Forgetting, guessing, and mastery: The Macready and Dayton models revisited and compared with a latent trait approach. *Journal of Educational Statistics,* 1978, *3,* 305-318.

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

van der Linden, W. J. Criterion-referenced measurement: Its main applications, problems, and findings. In W. J. van der Linden (Ed.), Aspects of criterion-referenced measurement. *Evaluation in Education: An International Review Series,* in press.

van der Linden, W. J. *The use of moment estimators for mixtures of two binomials with one known success parameter.* Submitted for publication, 1981.

Walsh, J. E. Approximate probability values for observed number of successes from statistically independent binomial events with unequal probabilities. *Sankhyá,* Series A, 1953, *15,* 281-290.

Walsh, J. E. Definition and use of generalized percentage points. *Sankhyá,* Series A, 1959, *21,* 281-288.

Walsh, J. E. Corrections to two papers concerned with binomial events. *Sankhyá,* Series A, 1963, *25,* 247.

Werts, E. W., Linn, R. L., & Jöreskog, K. A congeneric model for platonic true scores. *Educational and Psychological Measurement,* 1973, *33,* 311-318.

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox, *Achievement test items: Methods for study* (CSE Monograph No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977. (a)

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox, *Achievement test items: Methods for study* (CSE Monograph No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977. (b)

Wilcox, R. R. Achievement tests and latent structure models. *British Journal of Mathematical and Statistical Psychology,* 1979, *32,* 61-71. (a)

Wilcox, R. R. An alternative interpretation of three stability models. *Educational and Psychological Measurement,* 1979, *39,* 311-316. (b)

Wilcox, R. R., & Harris, C. W. On Emrick's "An evaluation model for mastery testing." *Journal of Educational Measurement,* 1977, *14,* 215-218.

Acknowledgments

I thank Gideon J. Mellenbergh for his helpful comments on the original version of this paper, Pieter Hoekstra for his computational assistance, and Paula Achterberg for typing the manuscript.

Author's Address

Wim J. van der Linden, Afdeling Toegepaste Onderwijskunde, Technische Hogeschool Twente, Postbus 217, 7500 AE Enschede, The Netherlands.