On the Efficiency of IRT Models When Applied to Different Sampling Designs

Martijn P. F. Berger
University of Twente

The problem of obtaining designs that result in the greatest precision of the parameter estimates is encountered in at least two situations in which item response theory (IRT) models are used. In so-called two-stage testing procedures, certain designs may be specified that match difficulty levels of test items with abilities of examinees. The advantage of such designs is that the variance of the estimated parameters can be controlled. In situations in which IRT models are applied to different groups, efficient multiple-matrix sampling designs are applicable. The choice of matrix sampling designs will also influence the variance of the estimated parameters. Heuristic arguments are given here to formulate the efficiency of a design in terms of an asymptotic generalized variance criterion, and a comparison is made of the efficiencies of several designs. It is shown that some designs may be found to be most efficient for the one- and two-parameter model, but not necessarily for the three-parameter model. *Index terms: efficiency, generalized variance, item response theory, optimal design.*

The notion of information about parameters in item response theory (IRT) models has led to several applications. In test construction and item selection, for example, Theunissen (1985), van der Linden (1987), and van der Linden and Boekkooi-Timminga (1989) used the information about ability (θ) parameters to obtain optimal item selection procedures and test designs. Samejima (1977) demonstrated how information as a function of θ can be used in tailored/adaptive testing. Lord (1974), Lord and Wingersky (1985), and Thissen and Wainer (1982), among others, used the asymptotic standard errors obtained from the inverse of the information on the parameters to compare the relative efficiency of tests, models, and designs. Vale (1986) applied sampling designs to minimize equating errors. Lord (1980) and Hambleton and Swaminathan (1985) described applications of information as a function of the IRT parameters in various fields of measurement.

Lord (1962) investigated precision of the estimation of population means for an item domain and demonstrated that for a fixed number of item-person confrontations, the mean performance of a population for an item domain is estimated most reliably when each item is taken by a different sample of persons. A similar result was obtained empirically by Pandey and Carlson (1976). Although this result has been used to stress the importance of multiple-matrix sampling designs, the question still remains whether these designs are also efficient to estimate IRT parameters.

As the sampling concept from survey analysis gained currency in educational measurement, the interest in alternative sampling designs increased. In large-scale assessment studies, it would save considerable classroom administration time if only a subset of items instead of a whole test is administered to the examinees. Whether such a selection of items would influence the efficiency of the item parameter estimates will depend on the assumed IRT model.

The efficiency of designs is considered here in terms of a generalized variance criterion connected with IRT parameters, and this criterion is used to compare the efficiencies of designs for the one-,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 3, September 1991, pp. 293-306

0146-6216/91/030293-14$1.95

two- and three-parameter models. The results of this comparison are relevant to constructors of item banks who need to estimate the item parameters efficiently. The use of information in IRT models is discussed first, however, and heuristic arguments are given to propose the generalized variance criterion.

Information in IRT Models

The notion of information in any statistical model is connected with the estimation of unknown parameters. The amount of information is defined (Kendall & Stuart, 1973, p. 10) as

$$\text{Inf} = \text{E}\left[\frac{\partial \ln(L)}{\partial \theta}\right]^2 \quad , \tag{1}$$

where $\ln(L)$ is the log of the likelihood function L of a parameter θ,
E is the expected value, and
∂ is the partial derivative sign.

The information about the parameters in IRT models is defined similarly. For example, consider the three-parameter logistic model, which gives the probability of a correct response to item i ($i = 1, \ldots, n$) as a function of the $\theta_j \in (-\infty, +\infty)$ for examinee j ($j = 1, \ldots, N$):

$$P_i(\theta_j) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta_j - b_i)]\}^{-1} \quad , \tag{2}$$

where $b_i \in (-\infty, +\infty)$ and $a_i \in (0, +\infty)$ are the item difficulty and discrimination parameters, respectively, and $c_i \in (0,1)$ is the guessing parameter. If consistency holds when all the parameters are estimated simultaneously and when the number of examinees and the number of items becomes large simultaneously, then it would be reasonable to represent the information by

$$\text{E}\left[\frac{\partial \ln(L)}{\partial \xi_p} \frac{\partial \ln(L)}{\partial \xi_q}\right] \quad \text{for p, q} = 1, 2, \ldots, \text{m} \quad . \tag{3}$$

To avoid indeterminacy of the model, two parameters must be fixed (i.e., $m = 3n + N - 2$). The likelihood is L and $\xi = (\xi_p) = (\boldsymbol{\mu}, \boldsymbol{\theta})$, where

$$\boldsymbol{\mu} = (a_1, b_1, c_1, a_2, b_2, c_2, \ldots, a_n, b_n, c_n) \tag{4}$$

and

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{N-2}) \quad . \tag{5}$$

Lord and Wingersky (1985) suggested gathering the information in the response data about the m parameters in the following partitioned matrix:

$$\mathbf{Inf}_3 = \left[\begin{array}{ccc|ccc} \mathbf{I}_1 & & 0 & \mathbf{K}_{11} & \mathbf{K}_{12} \cdots & \mathbf{K}_{1(N-2)} \\ & \mathbf{I}_2 & & \mathbf{K}_{21} & \mathbf{K}_{22} \cdots & \mathbf{K}_{2(N-2)} \\ & & \ddots & \vdots & & \vdots \\ 0 & & \mathbf{I}_n & \mathbf{K}_{n1} & \mathbf{K}_{n2} \cdots & \mathbf{K}_{n(N-2)} \\ \hline \mathbf{K}'_{11} & \mathbf{K}'_{21} \cdots & \mathbf{K}'_{n1} & \mathbf{J}_1 & & 0 \\ \mathbf{K}'_{12} & \mathbf{K}'_{22} \cdots & \mathbf{K}'_{n2} & & \mathbf{J}_2 & \\ \vdots & & \vdots & & & \ddots \\ \mathbf{K}'_{1(N-2)} & \mathbf{K}'_{2(N-2)} \cdots & \mathbf{K}'_{n(N-2)} & 0 & & \mathbf{J}_{(N-2)} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{K} \\ \hline \mathbf{K}' & \mathbf{J} \end{array}\right] . \quad (6)$$

The $3n \times 3n$ superdiagonal matrix $\mathbf{I}$ contains the 3×3 item information matrices $\mathbf{I}_1$ through $\mathbf{I}_n$ for n different items for the parameters a_i, b_i, and c_i, respectively. Note that the off-diagonal elements of $\mathbf{I}$ are 0, although the estimated covariances among estimated item parameters for different items may not be. The $(N - 2) \times (N - 2)$ diagonal matrix $\mathbf{J}$ contains Fisher's information $\mathbf{J}_1$ through $\mathbf{J}_{N-2}$ for θ_j, and $\mathbf{K}_{ij}$ is the 3×1 joint Fisher information vector for item i and person j. Thus the information in the data for the simultaneous estimation of the m parameters of the three-parameter logistic model are stored in the information matrix $\mathbf{Inf}_3$, and the asymptotic variances and covariances of the m estimated parameters can be obtained from $\mathrm{Cov}(\hat{\xi}) = \mathbf{Inf}_3^{-1}$.

The three-parameter model in Equation 2 will reduce to a two-parameter model if it is known that $c_i = 0$, and it will reduce to a one-parameter model if it is known that $a_i = 1$ and $c_i = 0$. However, a distinction should be made between a model with parameters having certain values and the information on these parameters needed for joint estimation. For example, if it is known that $c_i = 0$, c_i does not have to be estimated from a sample. Equation 2 will not only reduce to a two-parameter model, but the information on c_i and the joint information of c_i with the other parameters will not be needed. In addition, $\mathbf{Inf}_3$ will reduce to the information matrix $\mathbf{Inf}_2$ for the two-parameter model. If it is not known that $c_i = 0$, however, c_i will have to be estimated jointly with the other parameters, and all information in $\mathbf{Inf}_3$ will be needed.

It should also be noted that the information matrices $\mathbf{Inf}_1$ and $\mathbf{Inf}_2$ for the one- and two-parameter logistic model can be obtained from $\mathbf{Inf}_3$ by deleting the appropriate rows and columns of $\mathbf{I}$ and the corresponding rows of $\mathbf{K}$, and by adding a column to $\mathbf{K}$ and a diagonal element to $\mathbf{J}$ corresponding to examinee $N - 1$ for $\mathbf{Inf}_1$.

Two Different Approaches to Estimating Standard Errors

Two approaches have been employed to obtain the asymptotic standard errors of the estimated item parameters under the assumption that the examinees are not randomly sampled from a certain population. The difference between these two approaches is that the first approach (de Gruijter, 1984; Thissen & Wainer, 1982) assumes θ_j to be known, whereas the second approach (de Gruijter, 1985, 1988; Lord & Wingersky, 1985; Wingersky & Lord, 1985) does not assume that θs are known, but rather that they have to be estimated. It is argued here that the approach that assumes θ_j to be unknown is in fact implemented by using the information on the item parameters corrected for the joint information with the θs.

The first approach uses the fact that the maximum likelihood estimator $\hat{\boldsymbol{\mu}}_i$ of the triplet $\boldsymbol{\mu}_i = (a_i, b_i, c_i)$ is asymptotically normally distributed with covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\mu}}_i)$ (Kendall & Stuart, 1973, p. 59), which can be obtained from the inverse of the information matrix:

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_1 & & & \\ & \mathbf{I}_2 & & 0 \\ & & \ddots & \\ 0 & & & \mathbf{I}_n \end{bmatrix} \quad . \tag{7}$$

The elements of the 3 × 3 information matrices $\mathbf{I}_1$ through $\mathbf{I}_n$ can be obtained from Lord (1980, p. 191), and the standard errors of the estimated parameters can be obtained from the diagonal elements of $\text{Cov}(\hat{\boldsymbol{\mu}}) = \mathbf{I}^{-1}$.

The second approach is based on the full information matrix $\mathbf{Inf}_3$ (or $\mathbf{Inf}_1$ and $\mathbf{Inf}_2$ for the one- and two-parameter models, respectively), and assumes that the N θ_j parameters are unknown and must be estimated jointly with the item parameters. The covariance matrix of the estimated item parameters is now denoted by $\text{Cov}^*(\hat{\boldsymbol{\mu}})$ and is obtained from the leading $3n \times 3n$ matrix of $\mathbf{Inf}_3^{-1}$ (or the leading $n \times n$ matrix of $\mathbf{Inf}_1^{-1}$, or the leading $2n \times 2n$ matrix of $\mathbf{Inf}_2^{-1}$). The asymptotic standard errors of the estimated parameters are obtained from the diagonal elements of $\text{Cov}^*(\hat{\boldsymbol{\mu}})$. The asymptotic standard errors from these two approaches are not equal. See McLaughlin and Drasgow (1987) for a comparison of these standard errors.

The difference between the two approaches is explained below in terms of information and "partial" information. It is demonstrated how these quantities can be computed for any non-singular transformation of the parameters in an IRT model.

Information and "Partial" Information

A general result on partitioned matrices is given below that demonstrates that the use of the diagonal elements of $\text{Cov}^*(\hat{\boldsymbol{\mu}})$ is in fact based on the information on the item parameters corrected for the joint information with the θ parameters.

Let $\mathbf{H}$ be an $m \times m$ non-singular transformation matrix that is partitioned as $\mathbf{H} = (\mathbf{H}_1\ \mathbf{H}_2)$. The $m \times p$ matrix $\mathbf{H}_1$ and the $m \times (m - p)$ matrix $\mathbf{H}_2$ are of ranks $p \leq m$ and $(m - p)$, respectively. The information matrix $\mathbf{Inf}_3$ of m parameters can be transformed into a partitioned matrix containing information on two different sets of transformed parameters:

$$\mathbf{H}'\mathbf{Inf}_3\mathbf{H} = \left[\begin{array}{c|c} \mathbf{H}_1'\mathbf{Inf}_3\mathbf{H}_1 & \mathbf{H}_1'\mathbf{Inf}_3\mathbf{H}_2 \\ \hline \mathbf{H}_2'\mathbf{Inf}_3\mathbf{H}_1 & \mathbf{H}_2'\mathbf{Inf}_3\mathbf{H}_2 \end{array}\right] \quad . \tag{8}$$

Let $\mathbf{T}$ be a $p \times m$ matrix, such that $\mathbf{TH} = \mathbf{(Id0)}$. $\mathbf{Id}$ is an identity matrix and $\mathbf{0}$ is a zero matrix. The columns of $\mathbf{H}_1$ form the basis of the vector space spanned by the rows of $\mathbf{T}$, and $\mathbf{H}_2$ is a so-called deficiency matrix. If the information matrix $\mathbf{Inf}_3$ is symmetric and positive definite, the following equality holds (Lemma 1 of Khatri, 1966, p. 76-77):

$$\mathbf{Inf}_3^{-1} - \mathbf{Inf}_3^{-1}\mathbf{T}'(\mathbf{TInf}_3^{-1}\mathbf{T}')^{-1}\mathbf{TInf}_3^{-1} = \mathbf{H}_2(\mathbf{H}_2'\mathbf{Inf}_3\mathbf{H}_2)^{-1}\mathbf{H}_2' \quad . \tag{9}$$

From Equation 9, the "partial" information on the first set of transformed parameters can be obtained, given the information on the second set of transformed parameters. Pre- and post- multiplication of both sides of Equation 9 by $\mathbf{H}_1'\mathbf{Inf}_3$ and $\mathbf{Inf}_3\mathbf{H}_1$, respectively, and rearrangement of terms yields

$$\mathbf{H}_1'\mathbf{Inf}_3\mathbf{H}_1 - \mathbf{H}_1'\mathbf{Inf}_3\mathbf{H}_2(\mathbf{H}_2'\mathbf{Inf}_3\mathbf{H}_2)^{-1}\mathbf{H}_2'\mathbf{Inf}_3\mathbf{H}_1 = (\mathbf{TInf}_3^{-1}\mathbf{T}')^{-1} \quad . \tag{10}$$

This equation holds for different sampling designs and different parameterizations of the model. If different restrictions are used to take care of the indeterminacy of the model, for example, matrix **H** will contain the corresponding contrasts. De Gruijter (1988) uses such contrasts.

Although **H** and **T** may be chosen arbitrarily, as long as the condition $\mathbf{TH} = |\mathbf{Id0}|$ is satisfied, a very simple choice for **H** and **T** is

$$|\mathbf{H}_1\ \mathbf{H}_2| = \begin{bmatrix} \mathbf{Id} & \mathbf{0} \\ \mathbf{0} & \mathbf{Id} \end{bmatrix}, \quad \mathbf{T} = \mathbf{H}_1' \quad . \tag{11}$$

(3n × 3n; 3n × (N − 2))

This choice gives the same partitioning of $\mathbf{Inf}_3$ as given in Equation 6, and Equation 10 reduces to

$$\mathbf{I} - \mathbf{K}(\mathbf{J})^{-1}\mathbf{K}' = [\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})]^{-1} \quad . \tag{12}$$

This equation is comparable with Equations 8 and 14 from de Gruijter (1988) for the one- and two-parameter models, respectively.

Equations 10 and 12 indicate that when considering only the standard errors of the item parameters taken from the diagonal elements of $\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})$ (which is the $3n \times 3n$ leading matrix of the inverse of the full information matrix $\mathbf{Inf}_3$), the inverse of the "partial" information on the item parameters corrected for their joint information with the θ parameters is in fact being used. To simplify notation, the matrix **H** is dropped from the equations below.

An Optimality Criterion for Efficiency

It is clear from the above that there is nothing incorrect in using only the diagonal elements of the matrix $\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})$ as an indication of efficiency. Because the covariances of the estimated item parameters of an IRT model are not zero, however, it is preferable to use one of the criteria common in optimal design research that takes into account both the variances and covariances of the estimated item parameters and the estimation of the θ parameters.

Let the item parameters be of primary interest. The normal probability density of the estimator $\hat{\boldsymbol{\mu}}$ of the item parameters (Graybill, 1969) is then given by:

$$p(\hat{\boldsymbol{\mu}}) = (2\pi)^{-3n/2}|\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})|^{1/2}\exp\{-\tfrac{1}{2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'[\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})]^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\} \quad , \tag{13}$$

where $\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})$ is the leading $3n \times 3n$ matrix of $\mathrm{Cov}(\hat{\boldsymbol{\xi}})$ [for the one- and two-parameter models the matrix $\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})$ will be of order $n \times n$ and $2n \times 2n$, respectively].

Shannon (1948) proposed the following measure of uncertainty about the parameters, which is related to information theory and is associated with the probability density:

$$h[p(\hat{\boldsymbol{\mu}})] = -\mathrm{E}\{\ln[p(\hat{\boldsymbol{\mu}})]\} \quad . \tag{14}$$

Substitution of Equation 13 into 14 yields

$$\begin{aligned} h[p(\hat{\boldsymbol{\mu}})] &= -\mathrm{E}\left\{\tfrac{1}{2}\{-3n\ln(2\pi) - \ln|\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})| - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'[\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})]^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\}\right\} \\ &= \tfrac{1}{2}\left\{3n\ln(2\pi) + \ln|\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})| + \mathrm{Tr}\{[\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})]^{-1}\mathrm{E}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\}\right\} \\ &= \tfrac{1}{2}\left\{3n[\ln(2\pi) + 1] + \ln|\mathrm{Cov}^*(\hat{\boldsymbol{\mu}})|\right\} \quad . \end{aligned} \tag{15}$$

Thus, apart from the constant terms 1/2 and $3n[\ln(2\pi) + 1]/2$, the $\ln|\text{Cov}^*(\hat{\boldsymbol{\mu}})|$ reflects the amount of uncertainty about the parameters. Minimizing this function is equivalent to minimizing $|\text{Cov}^*(\hat{\boldsymbol{\mu}})|$. The criterion—the determinant of the covariance matrix of the estimated parameters—is often referred to as the generalized variance (Anderson, 1984) or the *D*-optimality criterion.

It must be emphasized that the question of efficient estimation of IRT parameters is more difficult than the efficient estimation of parameters in linear models. The main difficulty with IRT models is that the efficiency measure $|\text{Cov}^*(\hat{\boldsymbol{\mu}})|$ depends on the value of the unknown parameters, because the information matrix depends on these values. The values of the parameters to be estimated must be known before selecting an efficient design. To circumvent this difficulty, the criterion used in the next section is defined at the level of a parameter set $\boldsymbol{\mu}_i$ for an item i, and it is $|\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)|$, where $\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)$ is a main diagonal matrix of $\text{Cov}^*(\hat{\boldsymbol{\mu}})$. This makes it possible to express the criterion as a function of combinations of parameter values. Berger and van der Linden (1991) provide a review of other procedures to overcome this difficulty. Although $|\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)|$ takes into account both the variances and covariances of estimated item parameters and assumes unknown θs, it does not account for the covariances of estimated item parameters among the n items.

Finally, even though the generalized variance is a reasonable criterion to consider when dealing with the efficiency of a set of parameters, this criterion is model dependent. For example, if a design is investigated under the assumption that a model with m parameters is correct, but a model with $m - 1$ parameters is in fact more appropriate, then it is impossible to compare the efficiencies of these two models with this criterion.

Comparison of Designs

Suppose that an instructional program is designed to teach a number of skills to a population of examinees who can be grouped according to their ability to master a certain skill. Suppose also that there is an item domain containing a set of items that can be used to assess mastery of each individual skill, and that these items can be ranked according to their difficulties. It will generally be possible, for example, to make a distinction between examinees with high and low ability in solving certain mathematical problems. It is also generally possible to group the items of an arithmetic test in a design based on difficulty. A design might consist of administering the easy portion of the test to examinees with low ability, and the difficult portion to examinees with high ability. The efficiency of this design can be compared with another design in which all examinees take the entire test.

A design $D(\boldsymbol{\mu})$ can be characterized by the sampling procedure used to select examinees from a population and by the selection of items from an item domain. The items for the one-, two-, and three-parameter models are characterized by the parameter set $\boldsymbol{\mu}_i = \{b_i\}$, $\boldsymbol{\mu}_i = \{a_i,b_i\}$, and $\boldsymbol{\mu}_i = \{a_i,b_i,c_i\}$, respectively.

The relative efficiency of a particular design $D_2(\boldsymbol{\mu})$ compared to another design $D_1(\boldsymbol{\mu})$—both of which are used to estimate the same parameter set $\boldsymbol{\mu}_i$—can be defined as

$$\text{RE}_i[D_2(\boldsymbol{\mu}),D_1(\boldsymbol{\mu})] = \frac{|\text{Cov}_1^*(\hat{\boldsymbol{\mu}}_i)|}{|\text{Cov}_2^*(\hat{\boldsymbol{\mu}}_i)|} , \tag{16}$$

where $|\text{Cov}_1^*(\hat{\boldsymbol{\mu}}_i)|$ and $|\text{Cov}_2^*(\hat{\boldsymbol{\mu}}_i)|$ are the generalized variances of the estimated parameters for an item i in $D_1(\boldsymbol{\mu})$ and $D_2(\boldsymbol{\mu})$, respectively. If the relative efficiency is less than 1, the parameters of item i in Design $D_2(\boldsymbol{\mu})$ are estimated less efficiently than in Design $D_1(\boldsymbol{\mu})$; if the relative efficiency is

greater than 1, the item parameters in $D_2(\boldsymbol{\mu})$ are estimated more efficiently than in $D_1(\boldsymbol{\mu})$.

It is well known that the information on b_i for $c_i = 0$ is

$$\mathbf{I}_b = a_i^2\{\Sigma P_i(\theta_j)[1 - P_i(\theta_j)]\} \quad . \tag{17}$$

For fixed values of a_i, this information will be maximal when $P_i(\theta_j) = .5$ (i.e., when $\theta_j = b_i$). Thus, a sample of examinees with abilities equal to the item difficulty (i.e., with a standard deviation of abilities $SD_\theta = 0$) will give optimal information only when the difficulty parameter b_i is estimated (van der Linden, 1988). This will not generally hold true for the two- and three-parameter models, however, where a_i and c_i are also estimated. The relative efficiency measure in Equation 16 is used below to display the effect of the standard deviation of θ on the efficiency of the item parameters for the three IRT models.

Effect of Standard Deviation of θ on Information

The relative efficiencies for the one-, two-, and three-parameter models are given in Figure 1 as a function of SD_θ. (To restrict the range of the actual plotted values, the square roots of the relative efficiencies are displayed in the figures.) These relative efficiencies were computed for *one* item at a time with the following combinations of parameter values: $b_i = 0$; $c_i = 0$; and $a_i = .5$, 1.0, 1.5, and 2.0. Several designs were considered, with each design based on a sample of $N = 1{,}000$ normally distributed θs—all with the same mean θ, $M_\theta = b_i = 0$, but with different SD_θs ranging from .1 to 2.0. Thus the only difference between these designs was their different SD_θs. This made it possible to display the relationship between SD_θ and the efficiency of the item parameters. The relative efficiencies were computed directly from the parameters by using Equation 16 and by comparing the $|\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)|$ for each of these designs to the $|\text{Cov}_1^*(\hat{\boldsymbol{\mu}}_i)|$ for Design $D_1(\boldsymbol{\mu})$ with $M_\theta = 0$ and $SD_\theta = 2$. $\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)$ was a scalar for the one-parameter model, and $\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)$ was a matrix of orders 2×2 and 3×3, respectively, for the two- and three-parameter models.

As expected for the one-parameter model, decreasing the values of SD_θ will result in an increase in efficiency. For the two-parameter model, however, a decrease in SD_θ will *not* always lead to an increase in efficiency. For this model, the relative efficiency increases first, and eventually decreases as the SD_θ becomes smaller. This effect can be explained by the fact that the variance of the estimated discrimination parameter generally increases as SD_θ becomes smaller. Thus, the increasing variance of the discrimination parameter will dominate the outcome of the generalized variance as SD_θ becomes smaller. Note that a similar effect is also found in regression analysis, where the variance of the estimated slope is inversely related to the variance of the independent variable.

The pattern of relative efficiencies for the three-parameter model is the reverse of the pattern noted for the one-parameter model, and it differs from the pattern for the two-parameter model. It should be noted that, for the three-parameter model, the information on c_i and the joint information of c_i with the other two parameters is taken into account. For items of easy and average difficulty, the information on c_i is relatively small, but the joint information of c_i and b_i is relatively large, whereas the information on c_i is relatively large for difficult items. This causes the variances and covariances in the 3×3 matrix $\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)$ to differ from the corresponding variances and covariances in the 2×2 matrix $\text{Cov}^*(\hat{\boldsymbol{\mu}}_i)$ for the two-parameter model. Thissen and Wainer (1982) offer a further explanation for the differences in variance of the estimated parameters for the two- and three-parameter models. The most efficient design for the three-parameter model is one in which the θs have a relatively large standard deviation.

The relative efficiencies in Figure 1 suggest that efficiency will be gained by sampling examinees from subpopulations with relatively small differences in θ among examinees for the one- and two-

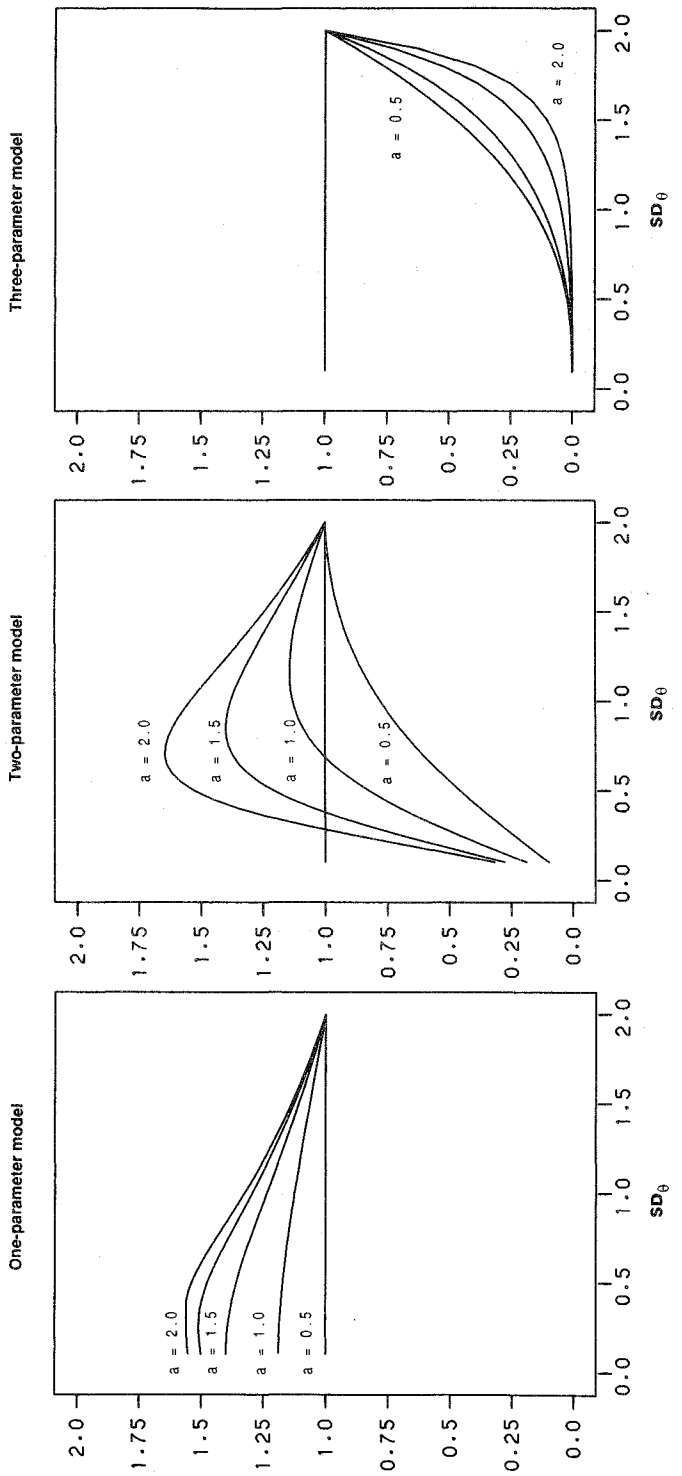


Figure 1
Relative Efficiency of Designs with SD_θ From .1 to 2.0

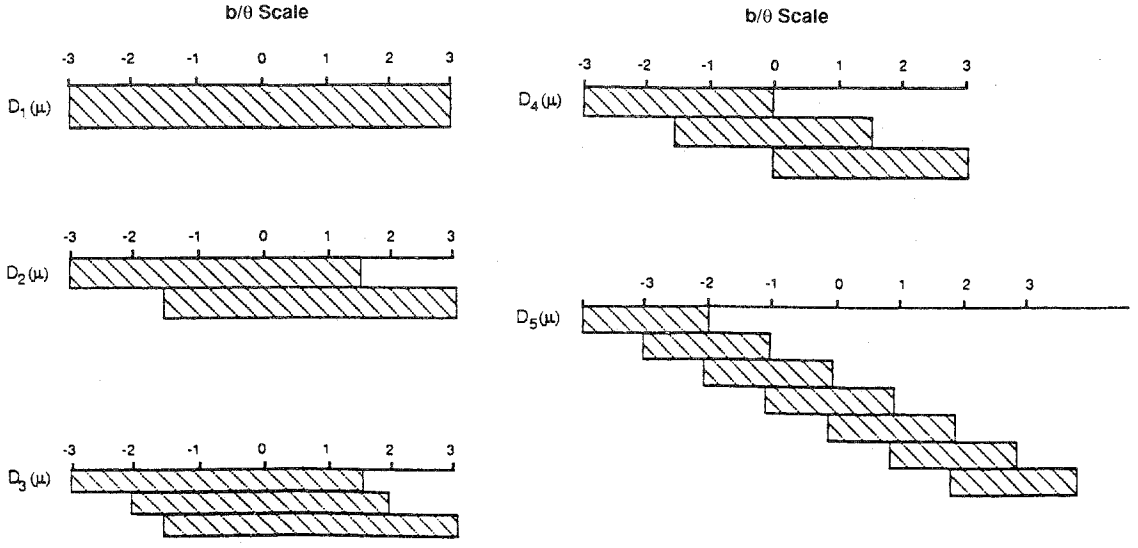
parameter models. For a test of items with difficulties b_i ranging from -3 to +3, for example, it appears to be more efficient to select a design with different samples of examinees with relatively small SD_θs and then administer items to each of these samples such that the item difficulties are approximately equal to the θs. This is in contrast to selecting a design in which items with b_i in the same range are administered to one sample of examinees with a large SD_θ. The comparison of designs below was made to investigate this conjecture.

Alternative Sampling Designs with Roughly Matched θs and Item Difficulties

Method. A limited comparison of designs was analyzed to illustrate the application of the proposed efficiency measure. Five designs are diagrammed in Figure 2 for a fixed number of examinee-item combinations ($N \times n$). It was assumed that the θs were uniformly distributed with a certain mean and range. The test consisted of $n = 7$ items, even though different groups took different items, as explained below. The computations were performed for all items having the same a_i value. This procedure was repeated for a_i values of .5, 1.0, 1.5, and 2.0. It was assumed that $c_i = 0$ for the three-parameter model. In addition, the efficiency measure given by Equation 16 was directly computed using item and θ parameter values. Results can differ when the relative efficiency measure is based on estimated item and θ parameters. Depending on the range of the θ scale employed, some items can be used as anchors for the actual estimation and scaling of item parameters.

1. Design $D_1(\mu)$: This design consisted of one sample of $N = 1{,}000$ examinees from a uniformly distributed population of θs ranging from -3 to +3, taking a test of $n = 7$ items with equally spaced difficulty parameters b_i ranging from -3 to +3. The total number of examinee-item combinations for this design and the four designs described below was approximately $N \times n = 7{,}000$.
2. Design $D_2(\mu)$: This design consisted of one group of 700 examinees with θs ranging from -3 to 1.5 taking the five easiest items in the test, and another group of 700 examinees with θs ranging from -1.5 to 3 taking the five most difficult items in the test.
3. Design $D_3(\mu)$: This design had three groups of 467 examinees each with θs ranging from -3 to

Figure 2
Five Designs for a Fixed Number of Examinee-Item Confrontations



1.5, -2 to 2, and -1.5 to 3, respectively. Each group took five items with difficulties within the range of the θ scale for that group. For example, Group 1 took the five items of the test with difficulties in the interval -3 to 1.5.

4. Design $D_4(\mu)$: This design consisted of three groups with 667 examinees each with θs ranging from -3 to 0, -1.5 to 1.5, and 0 to 3, respectively. Again, each group took three or four items with difficulties within the range of the θ scale of each group.
5. Design $D_5(\mu)$: The last design had seven different groups of 540 examinees each with θs ranging from -4 to -2, -3 to -1, -2 to 0, -1 to 1, 0 to 2, 1 to 3, and 2 to 4, respectively. Each group took one or two items with difficulties within the θ range of that group.

The matrix $\text{Cov}^*(\hat{\mu})$ for Designs $D_2(\mu)$ through $D_5(\mu)$ was computed for a test with $n = 7$ items having equally spaced b_i values ranging from -3 to +3, but with the same a_i value for all items. Thus the $\text{Cov}^*(\hat{\mu})$s were of orders 7 × 7, 14 × 14, and 21 × 21, respectively, for the one-, two-, and three-parameter models. The main diagonal matrix for each item was $\text{Cov}^*(\hat{\mu}_i)$. As previously noted, $\text{Cov}^*(\hat{\mu}_i)$ was a scalar for the one-parameter model. The relative efficiency for Designs 2 through 5 was obtained by dividing $|\text{Cov}_1^*(\hat{\mu}_i)|$ for $D_1(\mu)$ by $|\text{Cov}^*(\hat{\mu}_i)|$ for each of these designs.

Results. The relative efficiencies of $D_2(\mu)$ through $D_5(\mu)$ compared to Design $D_1(\mu)$ are given in Figures 3 and 4 for items with combinations of parameters $b_i \in (-3, +3)$, $a_i \in (.5, 2)$, and $c_i = 0$. Because Designs $D_2(\mu)$ through $D_5(\mu)$ are each related to $D_1(\mu)$, the results make it possible to compare these designs with each other.

It is clear from these figures that Design $D_4(\mu)$ is generally more efficient than $D_3(\mu)$ for the one- and two-parameter models, and Design $D_5(\mu)$ seems most efficient for tests with highly discriminating items. This can be explained by the smaller standard deviations of θ in each of the samples of $D_5(\mu)$. Design $D_4(\mu)$ seems most efficient for very easy items with large values for a_i for the three-parameter model, but Design $D_1(\mu)$ is superior for $b_i > 0$.

Computations were also performed for a 14-item test, and the results were quite comparable. Moreover, other sample sizes did not seriously affect the results, as long as the designs had the same number of examinee-item encounters.

Discussion

A generalized variance criterion to measure efficiency in IRT models was proposed and illustrated here. Although this criterion takes into account both the variances and covariances of maximum likelihood (ML) estimators of the item parameters and assumes that the θ parameters are fixed and unknown, it may not be optimal in all situations. Several other optimality criteria—defined as a function of the asymptotic variance-covariance matrix of ML estimators—have been proposed in the literature for optimal designs. One review of these criteria is given by Atkinson (1982). Each of these measures has advantages in specific situations and may be more or less sensitive to different scale restrictions of the parameters. The effect of scale restrictions in relation to the other optimality criteria needs to be studied more carefully.

Suppose that a test that measures certain skills is considered in which examinees are grouped according to their ability to master such skills. If the easy items from this test are administered to the examinees with lower ability, and the difficult items are administered to the examinees with higher ability, the application of the generalized variance criterion leads to the following recommendations for efficient estimation of the item parameters:

1. For the one- and two-parameter models, it would be more efficient to administer the easy half of a test to a sample of about .67N lower-ability examinees, the difficult half to another sample of about .67N higher-ability examinees, and the items of average difficulty to about .67N ex-

Figure 3
Relative Efficiency of Designs $D_2(\mu)$ and $D_3(\mu)$ Related to $D_1(\mu)$

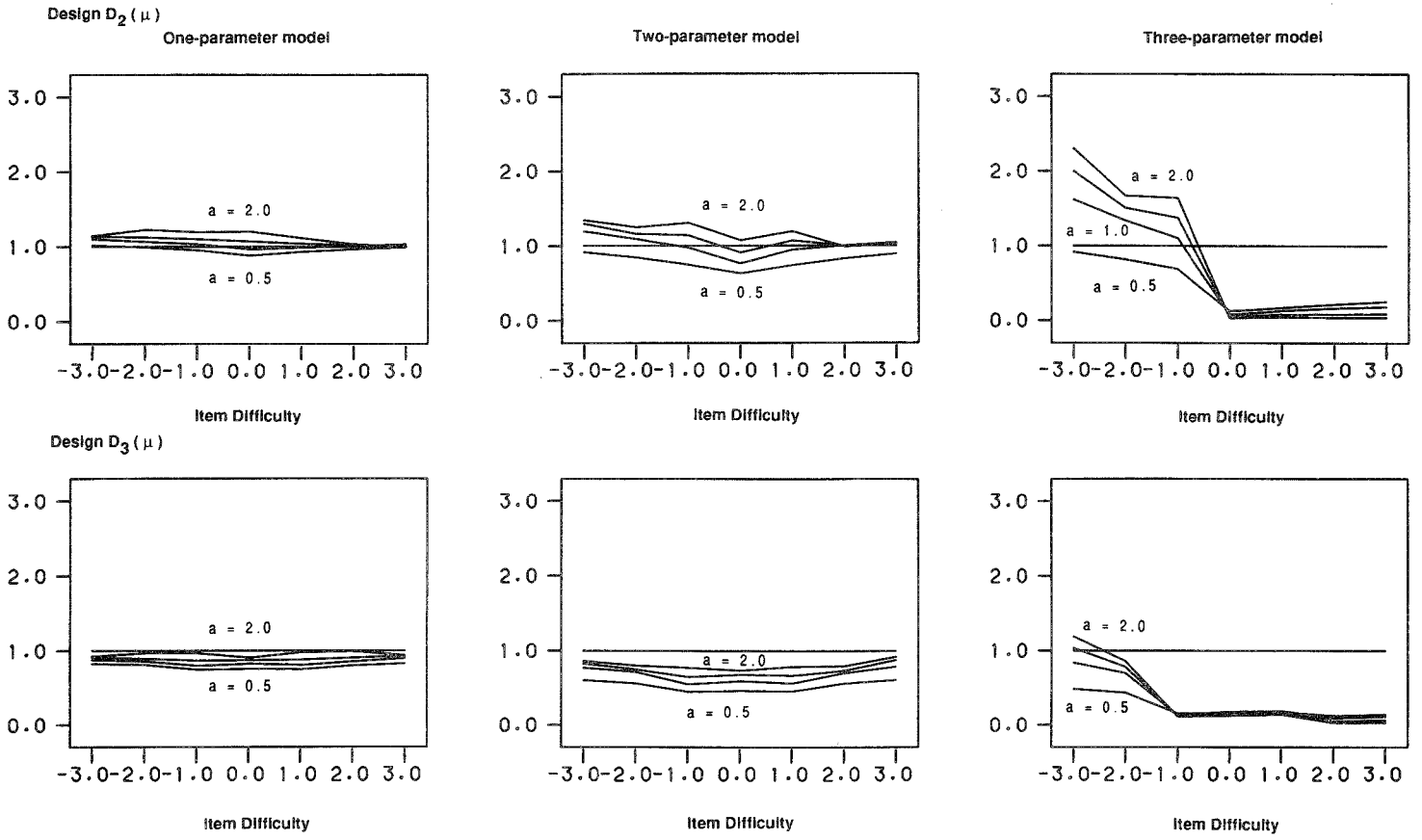
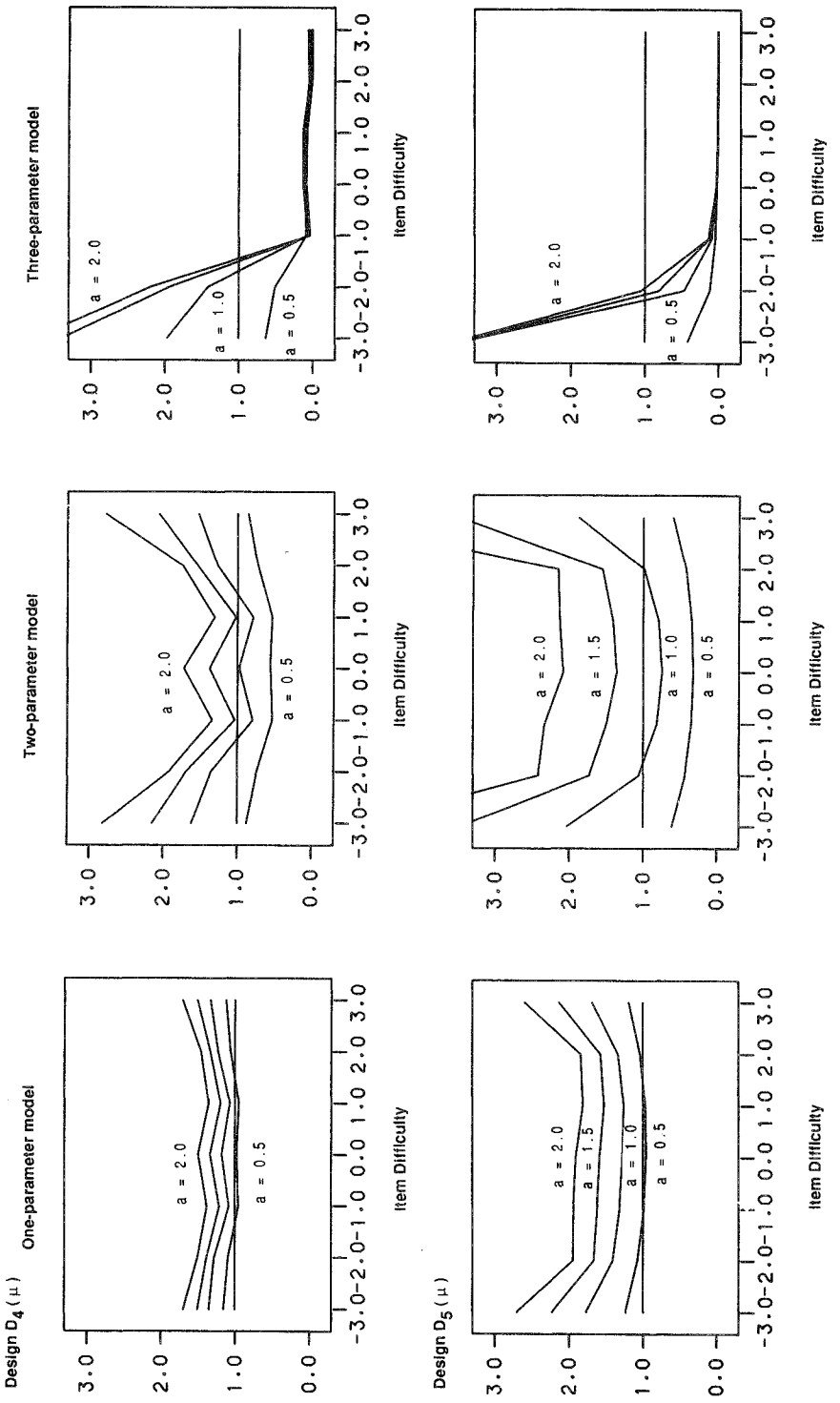


Figure 4
Relative Efficiency of Designs $D_4(\mu)$ and $D_5(\mu)$ Related to $D_1(\mu)$



aminees with average abilities—rather than to administer the entire test to one sample of size N. This will also be the case for the three-parameter model when easy, highly discriminating items are considered [i.e., Designs $D_1(\mu)$ versus $D_4(\mu)$]. The number of items to be administered per examinee in $D_4(\mu)$ is half as large as the number of items administered in $D_1(\mu)$. In practice, however, testing time may not be reduced to a similar degree, because it may take examinees longer to answer items of appropriate difficulty.

2. It would be better for all three IRT models to administer the entire test to a sample of N, rather than to administer approximately 70% of the test to three different samples of approximately $.47N$ examinees each in a design in which the easy section of the test would be administered to the lower-ability group and the difficult section would be administered to the higher-ability group [i.e., Designs $D_1(\mu)$ versus $D_3(\mu)$].
3. For the efficient estimation of parameters in the three-parameter model, it would be generally more efficient to use one sample of examinees with a relatively large SD_θ. Efficiency is gained only for easy items by administering them to a separate sample of lower-ability examinees.

The results reported here are based on a limited number of designs; however, they illustrate how the proposed efficiency measure can be applied. More research is needed to expand these conclusions to other designs. The importance of the proposed measure will be enhanced if its effectiveness is also evaluated with empirical datasets based on different designs.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Atkinson, A. C. (1982). Developments in the design of experiments. *International Statistical Review, 50,* 161–177.

Berger, M. P. F., & van der Linden, W. J. (1991). Optimality of sampling designs in item response theory models. In M. Wilson (Ed.), *Objective measurement: Theory into practice.* Norwood NJ: Ablex Publishing Company.

de Gruijter, D. N. M. (1984). A comment on some standard errors in item response theory. *Psychometrika, 49,* 269–272.

de Gruijter, D. N. M. (1985). A note on the asymptotic variance-covariance matrix of item parameter estimates in the Rasch model. *Psychometrika, 50,* 247–249.

de Gruijter, D. N. M. (1988). Standard errors of item parameter estimates in incomplete designs. *Applied Psychological Measurement, 12,* 109–116.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics.* Belmont CA: Wadsworth.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston MA: Kluwer-Nijhoff.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics, 18,* 75–86.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22,* 259–267.

Lord, F. M. (1974). The relative efficiency of two tests as a function of ability. *Psychometrika, 39,* 351–358.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 69–88). Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11,* 161–173.

Pandey, T. N., & Carlson, D. (1976). Assessing payoffs in the estimation of the mean using multiple matrix sampling designs. In D. N. M. de Gruijter & L. J. van der Kamp (Eds.), *Advances in psychological and educational measurement.* London: Wiley.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233–247.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411–420.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333–344.

van der Linden, W. J. (1987). *IRT-based test construction* (Research Rep. No. 87-2). Enschede, The Netherlands: University of Twente.

van der Linden, W. J. (1988). *Optimizing incomplete sampling designs for item response model parameters* (Research Rep. No. 88-5). Enschede, The Netherlands: University of Twente.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237–247.

Wingersky, M. S., & Lord, F. M. (1985). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Acknowledgments

The author expresses his appreciation to the two anonymous reviewers for their valuable comments, and to Nambury Raju for suggesting improvements in the text.

Author's Address

Send requests for reprints or further information to Martijn P. F. Berger, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.