PSYCHOMETRIKA—VOL. 57, NO. 4, 521–538
DECEMBER 1992

SEQUENTIAL SAMPLING DESIGNS FOR THE TWO-PARAMETER ITEM RESPONSE THEORY MODEL

MARTIJN P. F. BERGER

UNIVERSITY OF TWENTE, THE NETHERLANDS

In optimal design research, designs are optimized with respect to some statistical criterion under a certain model for the data. The ideas from optimal design research have spread into various fields of research, and recently have been adopted in test theory and applied to item response theory (IRT) models. In this paper a generalized variance criterion is used for sequential sampling in the two-parameter IRT model. Some general principles are offered to enable a researcher to select the best sampling design for the efficient estimation of item parameters.

Key words: sequential sampling design, efficiency, item response theory models, two-parameter logistic model, optimality criteria.

The problem of designing experiments has been encountered in various fields of research and several aspects have been studied in the statistical literature on optimal designs. Reviews of these developments are given by Atkinson (1982) and Steinberg and Hunter (1984), among others. Only in the past few years have the ideas from optimal design research been applied to test theory and item response theory (IRT) models. Berger and van der Linden (1992) give a review.

Designs and tests are generally optimized by maximizing information on the parameters of an IRT model. Van der Linden (1987) used information about the ability parameters to optimally design tests. Lord and Wingersky (1985), Thissen and Wainer (1982), and de Gruijter (1985, 1988) used asymptotic variances of the estimators of the item parameters to compare relative efficiencies of tests and models. Stocking (1990) uses information functions to find optimum ability levels. When IRT models are applied to different groups, matrix sampling designs can be used. The results of Lord (1962) and Pandey and Carlson (1976) stress the importance of multiple matrix sampling designs. Berger (1989, 1991) and van der Linden (1988) investigated the efficiency of some sampling designs for IRT models and Vale (1986) applied sampling designs to minimize equating errors.

In this paper the problem of optimal sampling of examinees for the two-parameter logistic model is considered. For the one-parameter logistic model it is well-known that the information on the difficulty parameter is maximal if all examinees have abilities equal to the (unknown) difficulty parameter (van der Linden, 1988). If abilities and item parameters are unknown, then estimates can be used to match abilities and item difficulties in so-called two-stage, or more generally, sequential testing procedures.

Some results will be given for the optimal sampling of examinees for the simultaneous estimation of item parameters in the two-parameter logistic model. These results are especially useful for large scale item calibration studies and for efficient estimation in two-stage and sequential design procedures. It should be noted that some of these problems can be solved by mathematical programming models. In this paper, however, easy and straightforward principles are given to enable practitioners to select an effi-

Requests for reprints should be sent to Martijn P. F. Berger, University of Twente, Department of Education, PO Box 217, 7500 AE Enschede, THE NETHERLANDS.

0033-3123/92/1200-91347$00.75/0

cient design for the two parameter model. These principles can of course also facilitate the choice of starting points in mathematical programming algorithms.

The Two-Parameter IRT Model

Consider the two-parameter logistic model, which gives the probability of a correct response ($U_{ij} = 1$) to item i ($i = 1, \ldots, n$) as a function of the ability parameter $\theta_j \in \mathbb{R}$ for examinee j ($j = 1, \ldots, N$):

$$P_i(\theta_j) = \{1 + \exp[-a_i(\theta_j - b_i)]\}^{-1}. \tag{1}$$

Item i is characterized by the pair of parameters $\{a_i, b_i\} \in \mathbb{R}^+ \times \mathbb{R}$, where $\mathbb{R}^+ \times \mathbb{R}$ is a two-dimensional rectangular set of positive real and real numbers, respectively. A whole test of n items is characterized by the pair of vectors $\{\mathbf{a}, \mathbf{b}\}$ containing a_i's and b_i's for each of the items.

A design for an IRT model is determined by the sampling procedure and can be denoted by the following vector of abilities $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]'$ and the corresponding vector of weights $\mathbf{W} = [w_1, w_2, \ldots, w_c]'$. The vector $\boldsymbol{\theta}$ consists of c distinct abilities θ_j and is also referred to as a vector of design points. It is assumed that $w_j \geq 0$ for $j = 1, \ldots, c$, $\sum_j w_j = N$, and $1 \leq c \leq N$. If $c = N$ and $w_j = 1$ for all c classes, then all examinees in the sample have different abilities. For $c = 1$ the sample consists of N examinees, all with the same abilities. The same design arises when $1 \leq c \leq N$ and all but one of the c weights w_j are equal to zero. The calibration of items will often take place without actually knowing the ability vector $\boldsymbol{\theta}$. In situations where it is not possible to exactly select examinees with certain θ_j- values, it will be assumed that the pair of vectors $\{\boldsymbol{\theta}, \mathbf{W}\}$ denotes a sampling design with abilities selected by means of a stratified random sample from a population of abilities. The class (strata) means are θ_j, and the mean of the population of abilities μ_θ is estimated by $(\sum_j w_j\theta_j)/\sum_j w_j$.

If the proportion of correct responses in each of the c classes or groups is given by $\mathbf{p} = [p_1, p_2, \ldots, p_c]'$, then the likelihood function based on $\boldsymbol{\theta}$ and $\mathbf{p}$ is

$$L = \prod_{j=1}^{c} P_i(\theta_j)^{w_j p_j}[1 - P_i(\theta_j)]^{w_j(1-p_j)}, \tag{2}$$

and the asymptotic efficiency of the maximum likelihood estimators of the two item parameters for a large sample of examinees is related to the Fisher information matrix:

$$J(a_i, b_i|\boldsymbol{\theta}, \mathbf{W}) = \sum_{j=1}^{c} w_j\{P_i(\theta_j)[1 - P_i(\theta_j)]\}\, \mathbf{X}\, \mathbf{X}', \tag{3}$$

where

$$\mathbf{X} = \begin{bmatrix} -(\theta_j - b_i) \\ a_i \end{bmatrix}.$$

The information matrix for a set of n items is a superdiagonal matrix $J(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}, \mathbf{W})$, with main diagonal matrices given by (3).

Each dichotomous response contains a certain amount of information about the parameters of interest. The amount of information in the sample depends on two factors, namely, the total number of responses and the variability of the responses in the sample. Because each response will cost money, a researcher will have to determine how many responses (i.e., how much information on the parameters) must be bought.

Too little information will prevent the researcher from obtaining good estimates, while too much information is a waste of time. The objective of using a sampling design for the estimation of item parameters is to maximize information on these parameters at a minimum cost. This can be done by selecting values for $\boldsymbol{\theta}$ and $\mathbf{W}$ that will maximize some function of the information contained by $J(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}, \mathbf{W})$ for the pair $\{\mathbf{a}, \mathbf{b}\}$. In the following section an optimality criterion for the maximization of information will be given.

Choice of An Optimality Criterion

In the literature on optimal designs, several optimality criteria have been proposed that are defined on the space of the information matrix. Atkinson (1982) reviews the three most familiar ones, namely, the D-optimality, A-optimality, and E-optimality criteria. Although each has certain advantages, we propose to apply the D-optimality or Determinant criterion, because this measure has some useful properties.

If $\hat{a}_i$ and $\hat{b}_i$ are the ML estimators of a_i and b_i and N is sufficiently large, then the likelihood region for a response $U_{ij} = u_{ij}$ is given by

$$\left[\{a_i, b_i\} : \sum_{j=1}^{c} w_j[\ln l_j(\hat{a}_i, \hat{b}_i) - \ln l_j(a_i, b_i)] \leq C\right], \tag{4}$$

where for each response u_{ij} under (1), $\ln l_j(a_i, b_i) = u_{ij}(a_i(\theta_j - b_i)) - \ln [\exp (a_i(\theta_j - b_i)) + 1]$, and C is a constant. This region can be approximated by the ellipse (Minkin, 1987):

$$[\{a_i, b_i\} : [a_i - \hat{a}_i, b_i - \hat{b}_i]J(\hat{a}_i, \hat{b}_i|\boldsymbol{\theta}, \mathbf{W})[a_i - \hat{a}_i, b_i - \hat{b}_i]' \leq 2C]. \tag{5}$$

For a given value of C, the volume of this ellipse is $V \propto \text{Det}\,[J(\hat{a}_i, \hat{b}_i|\boldsymbol{\theta}, \mathbf{W})^{-1/2}]$ (i.e., maximizing Det $[J(\hat{a}_i, \hat{b}_i|\boldsymbol{\theta}, \mathbf{W})]$ will minimize the volume of a confidence region in the parameter space). This criterion has already been proposed by Ward (1943) and is known as the generalized variance criterion (Anderson, 1984) of the D-optimality criterion (Kiefer, 1959), and is related to Shannon's (1948) information measure of uncertainty about the parameters (see Berger, 1991).

There is, however, a disadvantage connected with the use of this criterion. When the volume of an ellipse is minimized, this may result in an elongation in one direction and may lead to disturbing results. Another disadvantage is that the criterion depends on an adequate specification of the model. D-optimality criteria based on models with a different number of parameters are not comparable. It must be noted, however, that this problem also holds for the other well-known criteria.

One of the main advantages of the D-optimality criterion is that it is invariant under linear transformation of the parameter scale. Since it is related to the volume of a confidence region, it also has a natural interpretation. Moreover, some useful upper bounds of this criterion have been derived by Khan and Yazdi (1988). For these reasons the D-optimality criterion will be used to define optimality of sampling designs in the following paragraph.

The Fisher information matrix in (3) is a function of the parameters, and thus one will have to know the parameters before one can actually find a D-optimal design. One of the easiest solutions is to use the generalized variance criterion with some preliminary (ML) estimate of the parameters. Asymptotically this is justified, because the inverse of the expected information matrix is the asymptotic variance-covariance matrix of the parameter estimates. In small sample cases, however, the optimality of a

design will depend on the accuracy of the ML estimates. Generally, it will not be possible to locate one D-optimal design for all possible {**a**, **b**}. This is why the term "local optimality" has been introduced. Local optimality refers to optimality for a given pair of parameters {**a**, **b**}.

Analogous to a definition of a D-optimal design given by Khan and Yazdi (1988), the following definition of a locally D-optimal sampling design for a pair of parameter vectors {**a**, **b**} will be given.

Definition. A sampling design **D** {$\boldsymbol{\theta}^*$, **W***} with vector of ability parameters $\boldsymbol{\theta}^* = [\theta_1^*, \theta_2^*, \ldots, \theta_c^*]'$, $\boldsymbol{\theta}^* \in \mathbb{R}^c$, where $\mathbb{R}^c$ is a c-dimensional set of real numbers, and a vector of weights $\mathbf{W}^* = [w_1^*, w_2^*, \ldots, w_c^*]'$ is locally D-optimal if Det [J(**a**, **b**|$\boldsymbol{\theta}^*$, **W***)] $\geq$ Det [J(**a**, **b**|$\boldsymbol{\theta}$, **W**)], for a given pair {**a**, **b**} $\in \mathbb{R}^{n+} \times \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^c$ and **W**, with $\Sigma_j w_j = \Sigma_j w_j^*$.

An expression for the generalized variance criterion is given by

$$\text{Det } [J(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}, \mathbf{W})] = \prod_{i=1}^{n} \text{Det } [J(a_i, b_i|\boldsymbol{\theta}, \mathbf{W})], \tag{6}$$

where

$$\text{Det } [J(a_i, b_i|\boldsymbol{\theta}, \mathbf{W})] = \sum_{j=2}^{c} \sum_{j'=1}^{j-1} w_j w_{j'} (\theta_j - \theta_{j'})^2 G_j G_{j'},$$

and $G_j = a_i P_i(\theta_j)[1 - P_i(\theta_j)]$. Khan and Yazdi (1988) showed that when $P_i(\theta_j)$ is a logistic cumulative distribution function and $\boldsymbol{\theta} \in \mathbb{R}^c$ is a vector of abilities with weights **W**, the following inequality will hold:

$$\text{Det } [J(a_i, b_i|\boldsymbol{\theta}, \mathbf{W})] \leq \begin{cases} \text{Det } [J(a_i, b_i|\boldsymbol{\theta}^*, \mathbf{W}_1^*)] \text{ if } N \text{ is even,} \\ \text{Det } [J(a_i, b_i|\boldsymbol{\theta}^*, \mathbf{W}_2^*)] \text{ if } N \text{ is odd,} \end{cases} \tag{7}$$

where the design vector $\boldsymbol{\theta}^* = [-x, +x]'$ has weights $\mathbf{W}_1^* = [N/2, N/2]'$ if number N is even and $\mathbf{W}_2^* = [(N-1)/2, (N+1)/2]'$ if N is odd. It can be shown (Abdelbasit & Plankett, 1983; Khan & Yazdi, 1988) that for the two-parameter logistic model, $x = \theta - b_i = 1.5434/a_i$, and that the maximum obtainable value for Det [$J(a_i, b_i|\boldsymbol{\theta}, \mathbf{W})$] is $0.0501 \times N^2$.

From (7), the following two conclusions may be drawn: A design with equally weighted design points $\boldsymbol{\theta}^* = [-x, +x]'$ will have more information on a set of parameters $\{a_i, b_i\}$ than any other equally weighted $c = 2$ point design. This suggests that information on $\{a_i, b_i\}$ will become maximal when abilities are sampled symmetrically around b_i. Among all designs with points $\boldsymbol{\theta}^* = [-x, +x]'$, the most informative will be the one where examinees have abilities evenly distributed over the two points. This indicates that a bimodally distributed sample of abilities would be more informative for $\{a_i, b_i\}$ than a unimodal sample of abilities.

In the following section two types of designs will be discussed: two-stage designs where item parameters are unknown but can be estimated in the design procedure, and sequential designs which are very flexible and can easily be modified. First, however, the case of known item parameters will be considered.

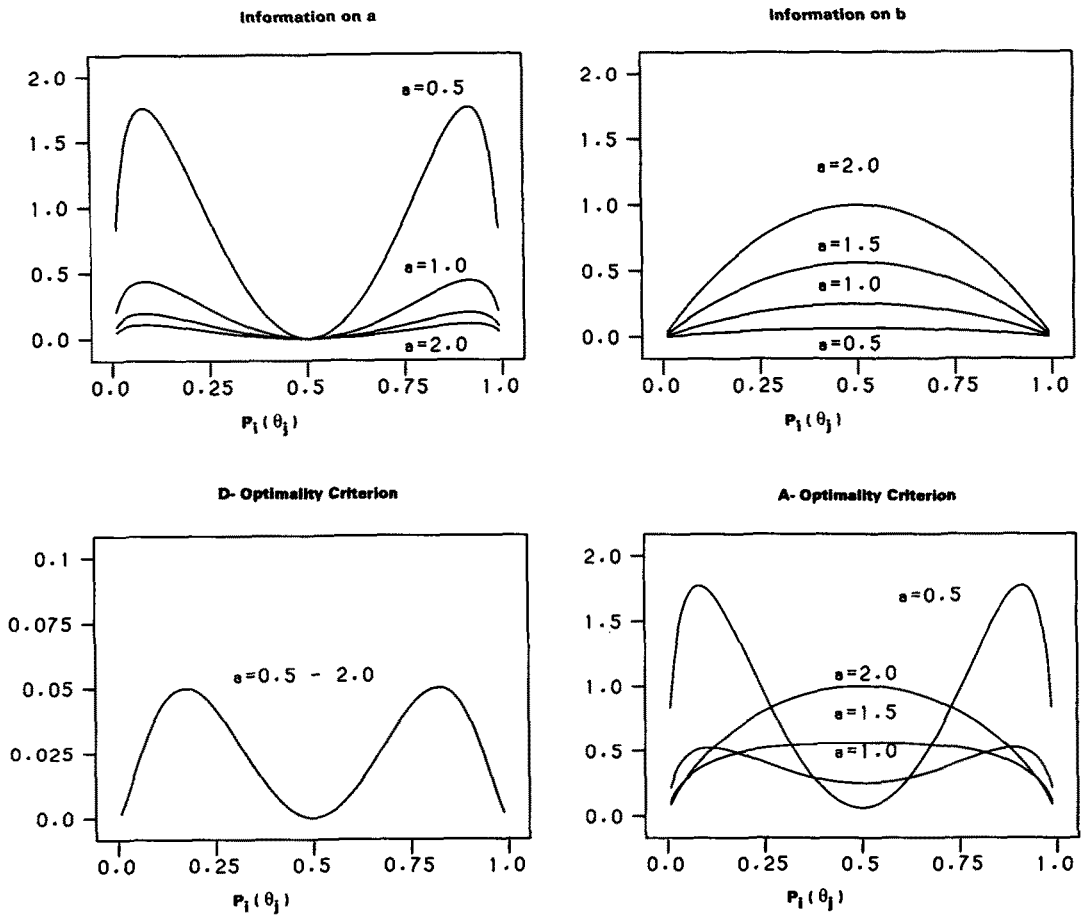

FIGURE 1.
D-optimality and A-optimality criteria based on information for the two-parameter IRT model with 2 design points.

Known Item Parameters

In this section some results are given for known item parameters. Although in most cases these item parameters are unknown, these results will help understand the underlying mechanisms.

In Figure 1 the information on the parameters a_i and b_i of the logistic model in (1) is given as a function of $P_i(\theta_j)$ for a two-point design $\boldsymbol{\theta} = [\theta_1, \theta_2]'$, with weights $\mathbf{W} = [N/2, N/2]'$. This design consists of a sample of N examinees with two distinct abilities θ_1 and θ_2, symmetric about b_i, such that $P_i(\theta_1) = 1 - P_i(\theta_2)$.

To restrict the range of the actual plotted values, the information on the parameters is divided by N. The computations show that the information on b_i is maximal for $P_i(\theta_1) = 1 - P_i(\theta_2) = 0.5$ and that the information on a_i is maximal for $P_i(\theta_1) = (1 - P_i(\theta_2)) = 0.08$. Figure 1 also shows that more information for estimating a_i is obtained when the value of a_i is low than when its value is high. An explanation for this phenomenon is given by Stocking (1990). As has been noted by Stocking, the abilities with maximal contributions to the information on a_i differ from the optimal abilities for b_i. The two abilities that contributed most to the information on a_i are those for which $P_i(\theta_j) = 0.08$ and $P_i(\theta_j) = 0.92$, respectively, while the ability that contributes most to the information on b_i is connected with $P_i(\theta_j) = 0.5$.

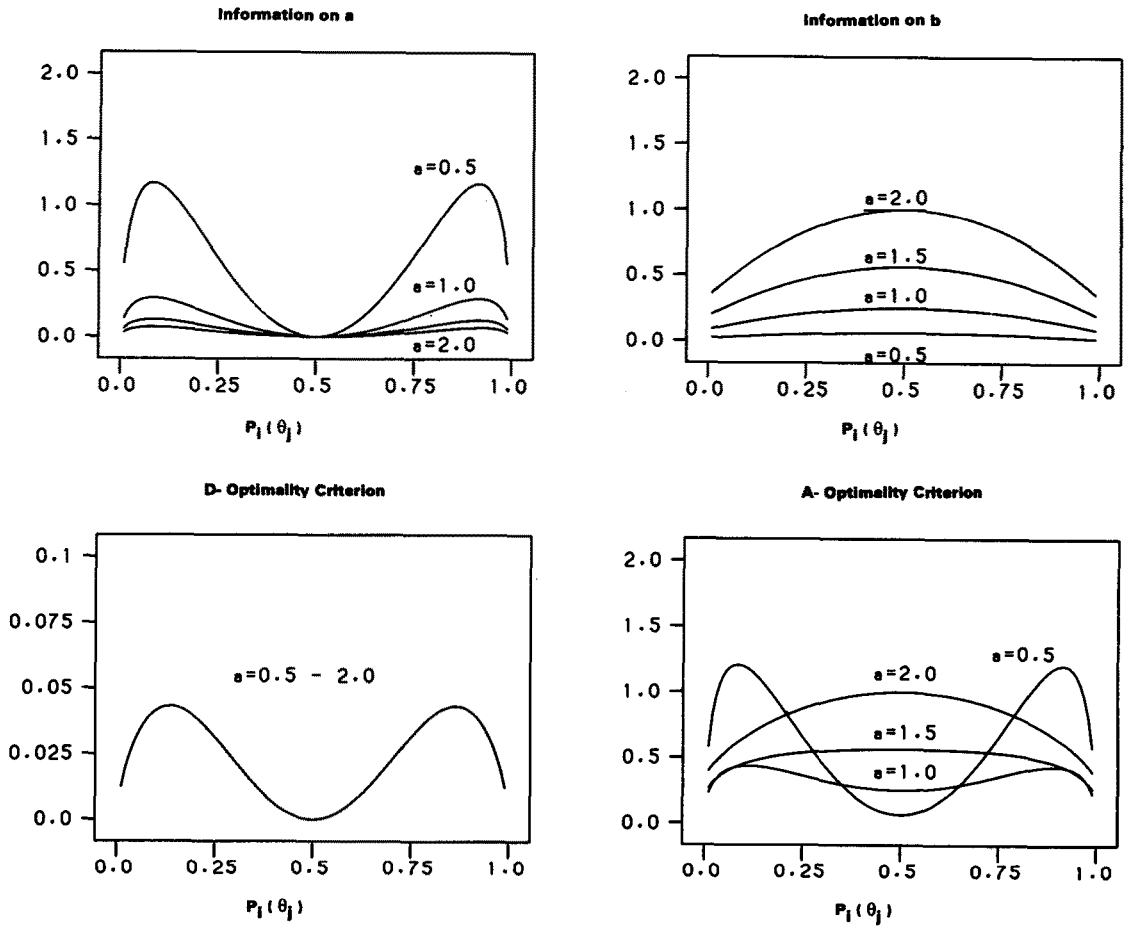

FIGURE 2.
D-optimality and A-optimality criteria based on information for the two-parameter IRT model with 3 design points.

The D-optimality criterion and the A-optimality or Trace criterion, which is the trace of the information matrix in (3) are also given in Figure 1. Since the information on a_i and b_i is divided by N, the value for the Trace criterion is also divided by N and the D-optimality criterion is divided by N^2. The D-optimality criterion has the same function value for different a_i values because this criterion is invariant under linear transformation of the parameter scale and for the logistic model a maximum value is obtained at $P_{\max} = P_i(\theta_1) = (1 - P_i(\theta_2)) = 0.824$. The A-optimality criterion, however, does not lead to such a result. Different values for a_i lead to different maximum values for this criterion.

Similar computations for a $c = 3$ point design $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]'$ with weights $\mathbf{W} = [N/3, N/3, N/3]'$, such that $P_i(\theta_1) = (1 - P_i(\theta_3))$ and $P_i(\theta_2) = 0.5$, are given in Figure 2. The maximum value for the D-optimality criterion is now located at $P_i(\theta_1) = (1 - P_i(\theta_3)) = 0.864$. The maximum D-optimality criterion value for this design is $0.0433 \times N^2$ and is somewhat smaller than that of the $c = 2$ point design. Figures 1 and 2 show that the D-optimality criterion for the $c = 3$ point design is only a little larger than that of the $c = 2$ point design when $P_i(\theta_j)$ approaches zero or one.

The results in these figures are symmetrical, in that $P_i(\theta_1) = 1 - P_i(\theta_2)$ for the $c = 2$ design, and $P_i(\theta_1) = 1 - P_i(\theta_3)$ for the $c = 3$ design. In concurrence with the

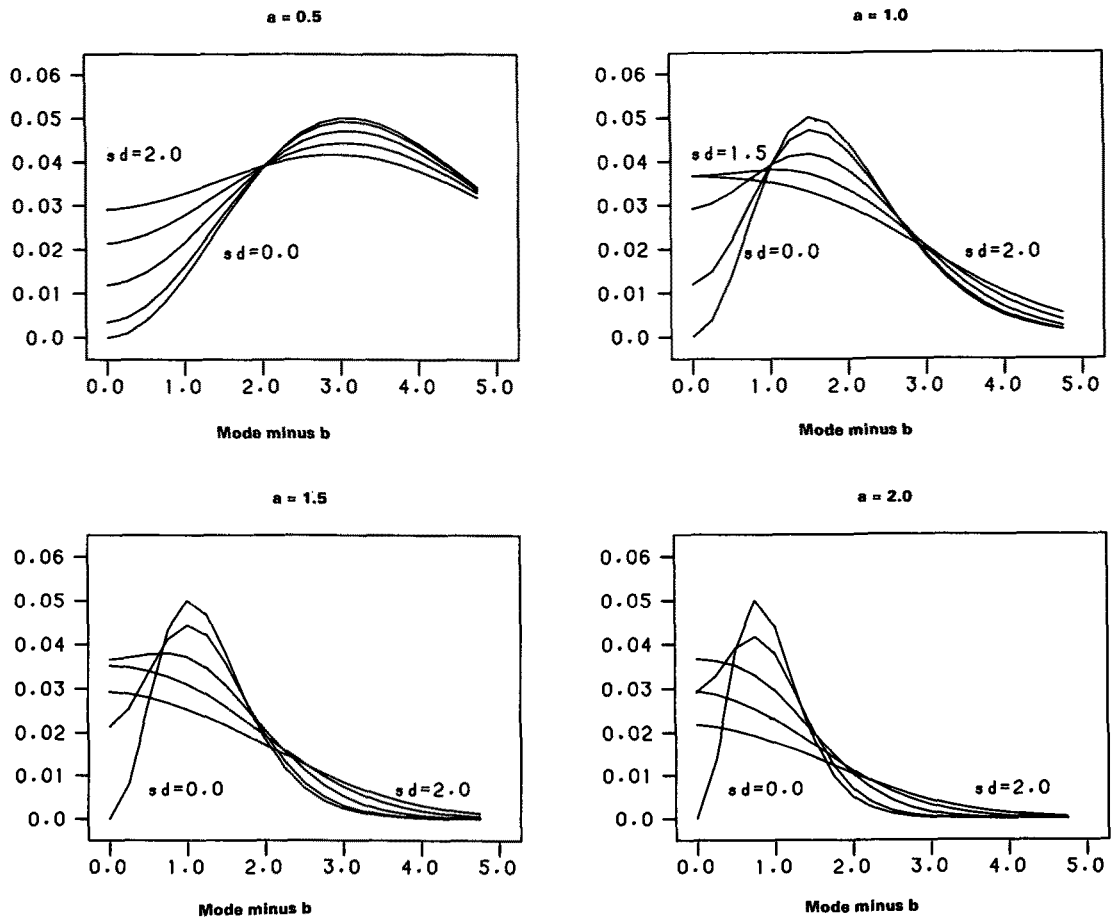

FIGURE 3.
D-optimality criterion for the two-parameter IRT model with a bimodal normal sample of abilities.

conclusions drawn from the previous section, these figures show that one should always choose design points that are located symmetrically around the item parameter b_i.

In most cases, however, the abilities will be unknown and the best thing to do is to sample abilities from a certain population. Since a design based on $c = 2$ equally weighted design points $[-x, +x]$ is the most optimal one, it seems reasonable to infer that a bimodal sample of abilities with modes $-x$ and $+x$ would approach such an optimal sampling design very well. In Figure 3 the results are given for samples selected from two normal distributions of abilities with modes located symmetrically around b_i (i.e., the differences between the two modes and b_i are equal). The five lines in each plot correspond to samples with $SD_\theta = 0.0, 0.5, 1.0, 1.5, 2.0$, respectively. The computations are given for $a_i = 0.5, 1.0, 1.5$, and 2.0. The D-optimality criterion is given as a function of the difference between each of the modes and b_i and is computed for different SD_θ values.

As already shown in Figure 1, the maximum value $0.0501 \times N^2$ is obtained for samples with $SD_\theta = 0$ and modes equal to $b_i - 1.5434/a_i$ and $b_i + 1.5434/a_i$, respectively. The results also show that as the value of a_i increases, the smaller the difference between the mode and b_i will have to be to approach maximal information. If we are prepared to accept about 80% of the maximal obtainable information, then

samples with $1 \leq SD_\theta \leq 2$, and a difference between the modes and b_i of about $1.5/a_i$ points, would be the most appropriate when a_i is 0.5 or 1.0. For large a_i values (i.e., 1.5 or 2.0), these samples would account for about 60% of the maximal achievable information.

Two-Stage Designs

The main purpose of using two-stage testing procedures has been to increase the precision of the ability estimates by matching item difficulty to the ability level of an examinee. Two-stage designs, however, can also be applied to increase the efficiency of item parameter estimates and to overcome the problem of unknown item parameters. If suitable initial estimates are available, these can be used and updated to increase efficiency in subsequent stages.

Suppose that N examinees are available and $\hat{a}_1$ and $\hat{b}_1$ are initial estimates of a_i and b_i. Then the following two-stage procedure which was proposed by Abdelbasit and Plankett (1983) can be used for a $c = 2$ point design where $P_{\max} = 0.824$ and $Q_{\max} = 1 - P_{\max}$.

First stage. The first stage consists of a design with vector $\boldsymbol{\theta}_f = [\theta_1, \theta_2]'$ and weights $\mathbf{W} = [N/4, N/4]'$, where:

$$\begin{aligned} \theta_1 &= \hat{b}_1 - [\ln (P_{\max}/Q_{\max})]/\hat{a}_1, \\ \theta_2 &= \hat{b}_1 - [\ln (P_{\max}/Q_{\max})]/\hat{a}_1. \end{aligned} \tag{8}$$

The design point θ_1 and θ_2 will maximize the criterion Det $[J(\hat{a}_1, \hat{b}_1|\boldsymbol{\theta}_f, \mathbf{W})]$ based on estimated parameters $\hat{a}_1$ and $\hat{b}_1$, and the accuracy of these initial estimates will determine how close the maximum value for the information on the true parameters is approached. The data obtained in the first stage can be used to update the parameter estimates. Let $\hat{a}_2$ and $\hat{b}_2$ be such updated estimates for use in the second stage.

Second stage. In the second stage a design is selected with vector $\boldsymbol{\theta}_s = [\theta_3, \theta_4]'$ and the same weights $\mathbf{W}$. The second stage design points are:

$$\begin{aligned} \theta_3 &= \hat{b}_2 - [\ln (P_{\max}/Q_{\max})]/\hat{a}_2, \\ \theta_4 &= \hat{b}_2 + [\ln (P_{\max}/Q_{\max})]/\hat{a}_2. \end{aligned} \tag{9}$$

These design points will maximize Det $[J(\hat{a}_2, \hat{b}_2|\boldsymbol{\theta}_s, \mathbf{W})]$. This maximization does not take into account the accuracy of the initial estimates, that is, accurate initial estimates do not have more weight than inaccurate initial estimates. The total information about the parameters for a sample of size N over the two stages is:

$$J(a_i, b_i|\boldsymbol{\theta}_t, \mathbf{W}_t) = J(a_i, b_i|\boldsymbol{\theta}_f, \mathbf{W}) + J(a_i, b_i|\boldsymbol{\theta}_s, \mathbf{W}), \tag{10}$$

where $\boldsymbol{\theta}_t = [\boldsymbol{\theta}_f, \boldsymbol{\theta}_s]'$ and $\mathbf{W}_t = [N/4, N/4, N/4, N/4]'$.

It should be emphasized that such a two-stage procedure maximizes the information on the estimated parameters in each step and that Det $[J(a_i, b_i|\boldsymbol{\theta}_t, \mathbf{W}_t)]$ is not maximized directly. But it may be assumed that Det $[J(a_i, b_i|\boldsymbol{\theta}_t, \mathbf{W}_t)]$ is a more accurate approximation of the maximal achievable criterion value than Det $[J(a_i, b_i|\boldsymbol{\theta}_f, \mathbf{W}^*)]$, where $W^* = [N/2, N/2]'$.

It should also be noted that although $J(a_i, b_i|\boldsymbol{\theta}_t, \mathbf{W}_t)$ can be used to construct an ellipse, there is no asymptotic theory to support the use of likelihood regions for sequential sampling inference.

The accuracy of Det $[J(a_i, b_i|\theta_t, \mathbf{W}_t)]$ will of course depend on the accuracy of both initial estimates $\hat{a}_1$ and $\hat{b}_1$. In Table 1 the relative efficiencies of a two-stage design are given for parameter values $a_i = 1.0$ and $b_i = 0.0$. The relative efficiency is obtained by dividing Det $[J(a_i, b_i|\theta_t, \mathbf{W}_t)]$ for the two-stage design by Det $[J(a_i, b_i|\theta_f, \mathbf{W}^*)]$ for a single-stage design. The results show that the two-stage design is more efficient than a one-stage design especially when the initial estimates differ from the actual parameter values. Note that when $\hat{b}_1 = b_i$ and $\hat{a}_1 = a_i$, the efficiency criterion is somewhat smaller than that of the single-stage design because of sampling fluctuations. The results in Table 1 also show that the advantage of a two-stage design over a single-stage design is greater than a_i is underestimated than when a_i is overestimated.

In Table 1 the criterion value Det $[J(a_i, b_i|\theta_t, \mathbf{W}_t)]$ is also related to the maximal achievable value for the $c = 2$ point design. These efficiencies are given within parentheses and show that the criterion is smaller when a_i is initially underestimated than when a_i is overestimated. This seems to indicate that overestimating a_i in the first stage will lead to a more optimal design than underestimating a_i.

The extension of this procedure to more stages is straightforward. The results in Table 1, however, indicate that an increase of the number of stages seems only worthwhile in terms of efficiency when the initial estimates are inaccurate. For small differences between the initial estimates and the parameters a two-stage procedure seems to approximate the maximum achievable efficiency very well. Such a two-stage procedure will become more complicated when a whole set of items is taken into account and both item and ability parameters are unknown and have to be estimated.

In the previous sections some mechanisms were explained to enable optimal estimation of the item parameters by selecting distinct samples of subjects. To verify whether these mechanisms also apply to situations where both item and ability parameters are unknown, the following simulation experiment was conducted.

Consider a set of achievement items that, for example, cover a 5th grade mathematics course. Let such a test consist of $n = 10$ items with unknown parameters $\mathbf{b} = [b_1, b_2, \ldots, b_n]'$ and $\mathbf{a} = [a_1, a_2, \ldots, a_n]'$, where $b_i \in [-3, +3]$ and $a_i \in [0, 2]$ for $i = 1, \ldots, n$. Item calibration usually takes place by administering the mathematics test to a random sample of pupils from a 5th grade population and by estimating the item parameter vectors **a** and **b** jointly. From the previous sections it can be inferred that it would be more efficient to estimate item parameters from data obtained from two samples of pupils. One sample, for example, contains 4th grade pupils for which the items are generally too difficult (i.e., they have a probability of answering the items correctly of approximately $P_{\max} = 1 - 0.824$), and a second sample of 6th grade pupils for which the items are generally too easy (i.e., with approximately $P_{\max} = 0.824$).

In the simulation experiment ten replications were performed. The parameter vectors **a** and **b** were generated for each replication from a uniform distribution and samples from distinct normal ability distributions were drawn. Since it is generally not very difficult to distinguish easy items from hard items, the $n = 10$ items were grouped into 5 easy and 5 hard items. Design A consists of three equally sized random samples of pupils. The first sample was drawn from a 4th grade population with mean ability $\mu_\theta = -2$ and variance $\sigma_\theta^2 = 1$. The second sample came from a 5th grade ability population with $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$, and the third sample came from a 6th grade population with $\mu_\theta = 2$ and $\sigma_\theta^2 = 1$. The first two samples take the five easy items and the last two samples take the five hard items. Thus each of the $n = 10$ items is administered to a total of N_A pupils from two equally sized samples. In design B one random sample of N_B pupils from a 5th grade, ability population with $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$ takes the whole test.

TABLE 1
The Relative Efficiency of a Two-Stage Design Compared to a Single-Stage Design with c=2 Design Points and b_i=0.0, a_i=1.0.

Initial Estimates for b_i	Initial Estimates for a_i										
	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50
-1.00	2.229 (0.676)	1.526 (0.762)	1.269 (0.819)	1.175 (0.853)	1.132 (0.854)	1.167 (0.872)	1.211 (0.869)	1.273 (0.862)	1.335 (0.844)	1.440 (0.845)	1.537 (0.834)
-0.80	1.991 (0.624)	1.516 (0.797)	1.159 (0.800)	1.087 (0.858)	1.101 (0.913)	1.068 (0.886)	1.099 (0.885)	1.183 (0.905)	1.236 (0.888)	1.266 (0.848)	1.414 (0.879)
-0.60	2.158 (0.693)	1.475 (0.806)	1.180 (0.858)	1.034 (0.870)	1.052 (0.940)	1.011 (0.911)	1.067 (0.938)	1.066 (0.897)	1.118 (0.888)	1.186 (0.881)	1.290 (0.893)
-0.40	2.079 (0.680)	1.465 (0.822)	1.189 (0.897)	1.013 (0.890)	1.026 (0.965)	1.009 (0.963)	1.008 (0.944)	1.062 (0.956)	1.070 (0.912)	1.134 (0.908)	1.226 (0.916)
-0.20	1.914 (0.632)	1.320 (0.753)	1.194 (0.920)	1.071 (0.967)	1.019 (0.989)	1.003 (0.991)	1.010 (0.983)	1.045 (0.979)	1.087 (0.967)	1.133 (0.949)	1.201 (0.940)
0.00	2.025 (0.671)	1.468 (0.842)	1.105 (0.850)	1.066 (0.971)	1.013 (0.954)	0.999 (1.000)	1.006 (0.992)	1.026 (0.975)	1.043 (0.942)	1.128 (0.958)	1.176 (0.935)
0.20	2.157 (0.712)	1.411 (0.805)	1.194 (0.920)	1.070 (0.966)	1.010 (0.980)	1.002 (0.991)	1.014 (0.987)	1.041 (0.975)	1.081 (0.962)	1.063 (0.890)	1.136 (0.890)
0.40	1.862 (0.609)	1.459 (0.819)	1.179 (0.889)	1.047 (0.921)	1.033 (0.972)	1.005 (0.960)	1.035 (0.970)	1.041 (0.937)	1.114 (0.949)	1.104 (0.884)	1.173 (0.877)
0.60	2.188 (0.703)	1.493 (0.816)	1.217 (0.885)	1.100 (0.924)	1.049 (0.937)	1.049 (0.945)	1.072 (0.942)	1.105 (0.930)	1.165 (0.925)	1.219 (0.906)	1.298 (0.899)
0.80	2.054 (0.644)	1.516 (0.797)	1.244 (0.859)	1.084 (0.855)	1.051 (0.905)	1.068 (0.887)	1.099 (0.884)	1.158 (0.886)	1.248 (0.897)	1.329 (0.890)	1.416 (0.880)
1.00	1.946 (0.590)	1.553 (0.775)	1.278 (0.824)	1.178 (0.855)	1.122 (0.846	1.178 (0.879	1.224 (0.878)	1.272 (0.862)	1.279 (0.809)	1.446 (0.849)	1.552 (0.842)

Note. The efficiency of a two-stage design related to the maximum achievable value is given in parentheses.

TABLE 2
The Cost Ratio N_B/N_A for a Single- and a Two-Stage Design

Replications		Single-Stage	Two-Stage	
			1st Stage	2nd Stage
1		1.38	1.06	1.32
2		1.49	1.06	1.39
3		1.42	1.29	1.39
4		1.68	0.96	1.47
5		1.46	1.10	1.42
6		1.55	1.29	1.47
7		1.33	1.08	1.28
8		1.48	1.44	1.35
9		1.53	1.13	1.34
10		1.35	0.94	1.28
	Mean	1.47	1.13	1.37
	SD	0.10	0.15	0.07

By means of the model in (1), data were generated and item parameters **a** and **b** estimated and rescaled for both designs. From the estimated parameters an estimate of Fisher information matrix can be obtained. If it can be assumed that the number of pupils taking an item is directly related to the cost of testing, the following ratio will give an indication of the costs related to the two designs:

$$\frac{N_B}{N_A} = \exp\left\{\frac{\log\left(\dfrac{\text{Det}\,[\hat{J}(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}_A, \mathbf{W}_A)]}{\text{Det}\,[\hat{J}(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}_B, \mathbf{W}_B)]}\right)}{2n}\right\}, \tag{11}$$

where $\hat{J}(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}_A, \mathbf{W}_A)$ is a $2n \times 2n$ super-diagonal matrix with estimated information on the item parameters. The cost ratios relating design A to B for a single-stage design procedure are presented in Table 2 under the column headed "Single-Stage".

These cost ratios show that for the single-stage procedure, about 1.5 times as many pupils will be needed in design B as in design A to obtain the same amount of information on the item parameters. This means that efficiency will be greater in design A than in design B.

Cost ratios for a two-stage design procedure are also presented in Table 2. The two-stage procedure differed from the single-stage procedure, in that the items are not first grouped into easy and difficult items. In the first stage the $n = 10$ items are randomly divided in two groups of five items each, and the information is estimated from the parameter estimates $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. In the second stage the items are grouped by means of their estimated $\hat{b}_i$'s into easy and hard items and the parameters estimated again by means of another sample. The results in Table 2 show that the cost ratios for

the first stage are approximately equal to one. Such a result, of course, can be expected. The cost ratios for the second stage, however, are about 1.4 (i.e., about 1.4 times the number of pupils will be needed in design B to obtain the same amount of information as in design A).

Such two-stage design procedures can easily be implemented but are also rather rigid. Not only does the test have to be divided into an easy and a difficult part, distinct populations or examinees must also be specified in advance. In the following section a more flexible procedure will be proposed.

Sequential Designs

One of the main problems in finding optimal designs is that it is often very inefficient to compute efficiencies for all combinations of weights with all possible values of $\boldsymbol{\theta}$. This is why the following simple procedure, which is an extension of an algorithm proposed by Wynn (1970), is offered. The procedure is based on a vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]'$, which contains all possible design points. The corresponding weights are $\mathbf{W} = [w_1, w_2, \ldots, w_c]'$ and the maximization of Det $[J(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}, \mathbf{W})]$ is only constrained by the sample size N. This means that the maximization procedure continues as long as $\sum_j w_j \leq N$.

Starting with an initial set of abilities $\boldsymbol{\theta}_0$ and weights $\mathbf{W}_0$, the procedure consists of a sequential addition of design points, such that in the $(k + 1)$-th step, the value θ_{k+1} with weight w_{k+1} is selected that has the largest possible value for

$$\prod_{i=1}^{n} \{\text{Det}\,[J(a_i, b_i|\boldsymbol{\theta}_k, \mathbf{W}_k) + w_{k+1}P_{i,k+1}Q_{i,k+1}\mathbf{X}_{k+1}\,\mathbf{X}'_{k+1}]\}. \tag{12}$$

$J(a_i, b_i|\boldsymbol{\theta}_k, \mathbf{W}_k)$ is the information matrix from the previous k steps and is based on the set of design points $\boldsymbol{\theta}_k$ and corresponding weights $\mathbf{W}_k$. $\mathbf{X}_{k+1} = [-(\theta_{k+1} - b_i), a_i]'$ and $P_{i,k+1} = [1 - Q_{i,k+1}]$, where $P_{i,k+1}$ is the probability of obtaining a correct response for θ_{k+1}. Wynn (1970, Theorem 1) showed that such a stepwise maximization will approximately lead to an optimal criterion value under the condition that for each step the design is admissible (i.e., for each step, $\{J(a_i, b_i|\boldsymbol{\theta}_k, \mathbf{W}_k)\}^{-1}$ is nonsingular). It can also be shown (see Wynn) that this procedure is equivalent to maximizing a variance function of the expected responses for each θ_{k+1}.

It must be emphasized that the sequence of selected design points is not unique and that several alternative designs may result in the same maximal criterion value. Thus, this procedure may lead to alternative solutions. In most cases, however, one is only interested in obtaining an optimal design as fast as possible and one is generally not interested in knowing whether such a design is unique.

For those cases where item and ability parameters are unknown, the procedure can be modified by replacing the parameters in (12) by their estimates. The information on the parameters can thus be updated in each step by newly estimated parameters. It must be noted, however, that in this case the stability of the procedure will be lower.

The procedure can also be applied to those cases where a sample of examinees is already available and one wants to increase efficiency by adding as few examinees as possible. Finally it should be mentioned that although usually for each step $w_k = 1$ (i.e., one design point is added in each step), it is possible to speed up the procedure by increasing the weights w_k of the design points for each step. In the following section this sequential procedure will be applied to some sets of real test items.

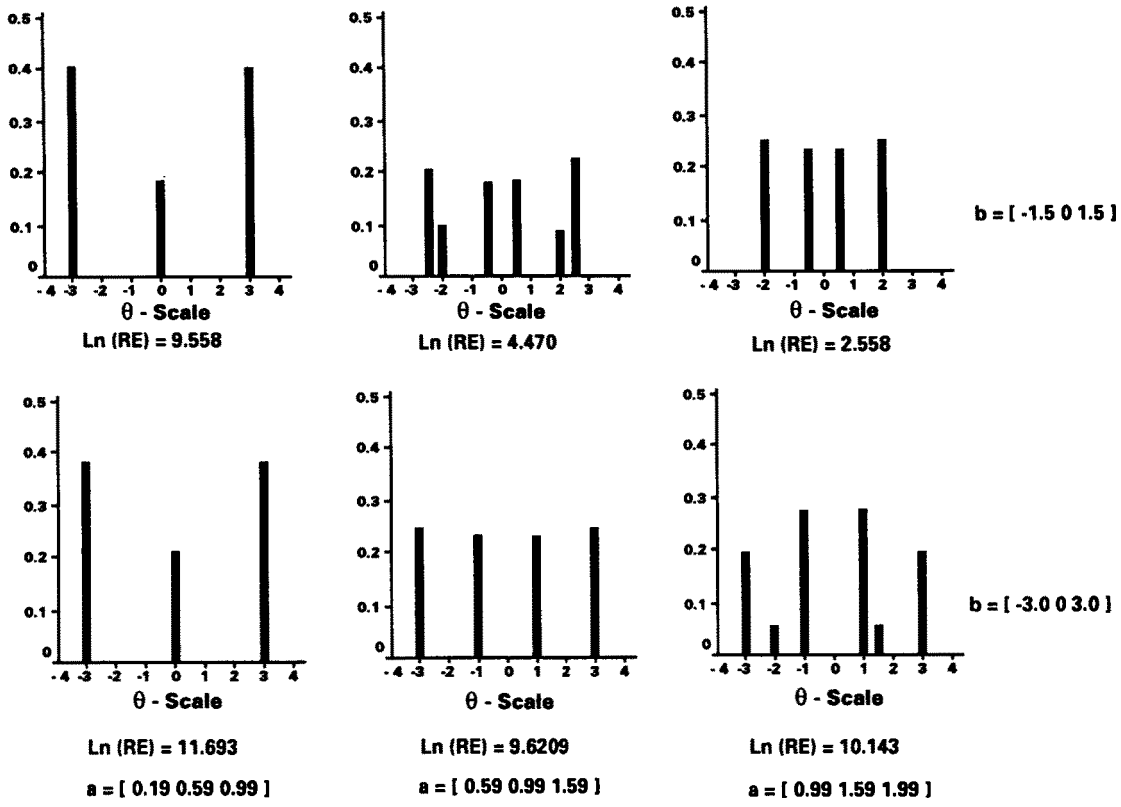

FIGURE 4.
Locally D-optimal sampling designs for typical sets of test items.

Examples

To illustrate the sequential design procedure, a typical set of test items given by Hambleton and Swaminathan (1985, p. 231) is analyzed. The set consists of two identical subsets having nine items each with the following combinations of parameters:

Items:	1	2	3	4	5	6	7	8	9
b	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
a	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3

To examine possible effects of variation in item parameter values, the following six combinations of parameters (i.e., six tests), including those given by Hambleton and Swaminathan (1985), are considered: $<b_1 = -1.5, b_2 = 0.0, b_3 = 1.5>$, $<b_1 = -3.0, b_2 = 0.0, b_3 = 3.0>$, and $<a_1 = 0.19, a_2 = 0.59, a_3 = 0.99>$, $<a_1 = 0.59, a_2 = 0.99, a_3 = 1.59>$, and $<a_1 = 0.99, a_2 = 1.59, a_3 = 1.99>$, respectively. Note that in many cases the logistic model in (1) is applied with a scaling factor $D = 1.7$. In those cases the a_i parameters should be lowered by this factor. The ranges of parameters estimated here will span most situations encountered in practice.

The procedure is based on a design vector $\boldsymbol{\theta} = [-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3]$ and starts with a weight vector of ones. Starting the procedure with other weights will generally not lead to different results for $N = 1000$. Smaller sample sizes, however, may show some different results.

The sequential design procedure is applied to each combination of item parameters and the resulting probability mass functions are given in Figure 4. For a test with low

discriminating items $<a_1 = 0.19, a_2 = 0.59, a_3 = 0.99>$, a locally optimal sampling design would have about 40% of the sample located at ability levels $\theta_1 = 3$ and $\theta_3 = -3$ and about 20% of the sample located at $\theta_2 = 0$. The results in Figure 4 also show that for higher discriminating items, the best design can be obtained by selecting abilities in such a way that the whole range for the difficulty parameter is divided into nearly equal pieces. For example, when $<b_1 = -3.0, b_2 = 0.0, b_3 = 3.0>$ and $<a_1 = 0.59, a_2 = 0.99, a_3 = 1.59>$, a locally optimal sampling design will be a design where about 25% of the sample has abilities equal to $\theta_1 = -3.0$, $\theta_2 = -1.0$, $\theta_3 = 1.0$, and $\theta_4 = 3.0$, respectively. The relative efficiencies (RE) for each of these optimal designs are also computed by comparing the Det $[J(\mathbf{a}, \mathbf{b}|\boldsymbol{\theta}, \mathbf{W})]$ of each design to the criterion based on a sample with normally $N(0, 1)$ distributed abilities. In Figure 4 the logarithms of the RE's are given. The results show that the relative efficiency is highest for tests with low discriminating items and a large range in difficulties. This indicates, that compared to a sample of normally distributed abilities the largest improvement of efficiency can be obtained for tests with hard and easy items having low discrimination parameter values. Substitution of these REs into (11) will show that these optimal designs lead to a considerable reduction of the sample size compared to samples with $N(0, 1)$ distributed abilities. For the six optimal designs in Figure 4 the percentage of examinees needed to obtain the same efficiency in estimating the item parameters as a sample with $N(0, 1)$ distributed abilities ranges from 52% to 87%.

To also provide somewhat more realistic examples, three different real tests were considered. The first test was the $n = 15$ TOEFL test given earlier by Wingersky and Lord (1984, p. 351–352). The estimated parameters were used as true parameters and are given in Table 3. The second test was a SAT test with $n = 15$ items. The estimated item parameters were based on responses of 2250 whites and given by Lord (1980, p. 221). These estimated parameters are also treated as true parameters. The last test has also been considered by McLaughlin and Drasgow (1987, p. 166). This test consisted of $n = 50$ items and the items parameters are also given in Table 3. It will be assumed that all these parameters come from model (1), without a scaling factor. As mentioned above, the inclusion of a scaling factor $D = 1.7$ means that the a_i parameters in Table 3 should be lowered by this factor.

For each of these tests the sequential design procedure was based on the same design vector $\boldsymbol{\theta}$ and started with the same weight vector as above. The probability mass functions for the locally optimal designs are presented in Figure 5. From Figure 5 it can be inferred that for all three tests an optimal design would be a design with a trimodal sample of abilities. Two modes are located at the extreme points of the difficulty (ability) scale and the third mode is located near the center of the difficulty (ability) scale. This result is in concurrence with that of the typical test with low discriminating items in Figure 4, because these tests mainly had low or moderately discriminating items. The logarithms of the REs of these designs, compared to a $N(0, 1)$ distributed sample of abilities, are also given in Figure 5, and give an indication of the efficiency of these designs. Substitution of the REs into (11) will show, for example, that the optimal design in Figure 5 for the test discussed by McLaughlin and Drasgow (1987) will reach the same efficiency in estimating the item parameters as a sample with $N(0, 1)$ distributed abilities with only 75% of the number of examinees. For the TOEFL test (Lord & Wingersky, 1985) and the SAT test (Lord, 1980) this percentage will be about 80% and 66%, respectively.

TABLE 3

Three Real Tests

Wingersky & Lord (1984) Data			Lord (1980) Data			McLaughlin & Drasgow (1987) Data					
Item	a_i	b_i	Item	a_i	b_i	Item	a_i	b_i	Item	a_i	b_i
1	0.99	-2.01	1	0.87	-1.5	1	1.1	-0.7	21	1.1	1.2
2	0.35	-1.61	2	0.28	-3.3	2	0.7	-0.6	22	1.2	1.1
3	1.38	-1.09	3	0.63	-1.1	3	0.4	0.1	23	1.3	0.2
4	0.78	-0.77	4	0.85	0.1	4	0.9	0.9	24	1.3	0.2
5	0.42	-0.67	5	0.35	-1.9	5	1.2	0.7	25	0.5	-0.8
6	0.92	-0.34	6	0.82	-0.4	6	1.6	1.1	26	0.7	0.5
7	0.92	-0.15	7	0.56	-0.6	7	1.6	1.1	27	0.7	0.5
8	1.06	0.00	8	0.50	1.2	8	1.6	-0.1	28	0.4	-0.4
9	1.34	0.11	9	1.43	0.5	9	1.2	0.5	29	0.4	-0.4
10	1.54	0.26	10	1.09	0.7	10	2.0	1.6	30	1.2	-0.5
11	0.87	0.46	11	1.64	1.7	11	1.0	1.6	31	0.7	-1.0
12	0.62	0.57	12	0.49	1.9	12	1.5	1.7	32	0.2	-0.2
13	1.09	0.68	13	1.63	1.7	13	1.0	0.7	33	0.7	-0.2
14	1.39	0.90	14	1.27	2.6	14	1.1	2.0	34	0.5	0.0
15	1.50	1.16	15	0.68	3.4	15	1.1	2.4	35	0.9	0.5
						16	2.0	1.4	36	1.1	1.4
						17	1.7	1.3	37	1.2	-0.6
						18	0.5	-0.6	38	1.2	-0.6
						19	0.9	1.6	39	0.6	-0.5
						20	1.3	0.4	40	1.6	0.3
									41	1.1	0.0
									42	1.5	2.0
									43	1.9	1.9
									44	0.9	-0.5
									45	0.7	-0.5
									46	1.4	1.6
									47	1.4	1.6
									48	1.0	1.7
									49	1.2	1.1
									50	1.2	1.1

Discussion

In this paper the idea of selecting a sampling design for the optimal estimation of item parameters in the two-parameter IRT model was explored. Since the item parameters are usually estimated jointly, it is better to consider an optimality criterion that takes into account both the information on the discrimination and difficulty parameter and the joint information on the two parameters, than to just examine the information on each of these parameters separately.

A generalized variance or D-optimality criterion was used to examine this problem.

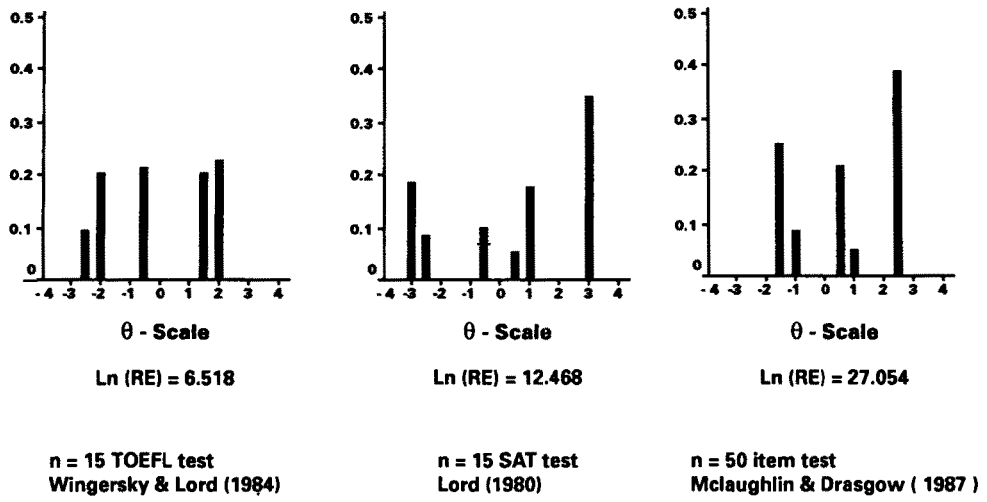

FIGURE 5.
Locally D-optimal sampling designs for two tests with $n = 15$ items (Lord, 1980; Wingersky & Lord, 1984) and an $n = 50$ item test (McLaughlin & Drasgow, 1987).

Although this criterion has several nice properties it depends on the unknown parameters and an optimal design selected by means of this criterion will also depend on unknown parameters. Several procedures to get around this problem have been suggested (see Berger & van der Linden, 1992). In this study the problem of unknown item parameters was handled by using two-stage designs and sequential designs, and the problem of not knowing the ability parameters was handled by drawing samples of examinees from distinct ability populations. The presented results showed that two-stage and sequential design procedures will lead to a considerable increase of efficiency in estimating the item parameters.

The result that a more uniform shape of the ability distribution will reduce the standard errors of the item parameters has been presented by Wingersky and Lord (1984) and was further supported by Stocking (1990). She argues strongly for the use of calibration samples of which the true ability distribution is close to uniform or even bimodal. Some of the results from this study concur with these conclusions. The presented two-stage design procedure makes use of the notion that an optimal criterion value is obtained when the sample has examinees with abilities located symmetrically around the item difficulty. The proposed sequential design procedure, however, leads to locally D-optimal sampling designs that are generally based on a trimodal sample of abilities. From the examples in this study, it may be inferred that for a whole test of items with distinct difficulties and low discrimination parameters, the most efficient sample would be a trimodal sample of abilities with two modes located at the extremes of the difficulty (ability) scale and the third mode located near the center of the scale. For a test with highly discriminating items it would be better to sample abilities uniformly from the whole ability scale. Although the presented locally optimal sampling designs only hold for the specific sets of items, this conclusion seems to apply to a variety of tests. The proposed sequential design procedure can easily be applied to other sets of test items. If it is not possible to provide a rough initial assessment of the item parameters, the procedure can be modified to provide sequential estimates of the item parameters. Since this will affect the stability of the procedure and possible bias of the estimates may occur, further research on the sequential estimation of item parameters will be needed.

The proposed procedures are applied to a two-parameter IRT model. Although the mechanisms can also be applied to a three-parameter model, the results may become quite different and the sequential design procedure may become less stable. Because of the well-known problems connected with the estimation of the lower asymptote c_i (Thissen & Wainer, 1982), and the possibility of multiple solutions to the likelihood function (Yen, Burket & Sykes, 1991), it is generally wise to avoid estimating c_i. In many cases where the three-parameter model is used, the lower asymptote is fixed in advance and updated after estimating the other item parameters. In these cases it is not necessary to incorporate the information on c_i and its joint information with the other item parameters in the sequential optimal design procedure, because c_i is not really estimated jointly with the other item parameters.

Finally it should be emphasized, that the advantage of using a more informative sampling design in terms of reducing the cost of testing, will depend heavily on the test situation. Even if it is possible to reduce the sample size by, for example, 33%, then this will only be worthwhile in practical testing situations when a substantial reduction in the cost of testing is also achieved.

References

Abdelbasit, K. M., & Plankett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association, 78*, 90–98.

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

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

Berger, M. P. F. (1989). *On the efficiency of IRT models when applied to different sampling designs* (Research Report 89-4). Enschede: University of Twente, Department of Education.

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement, 15*, 293–306.

Berger, M. P. F., & van der Linden, W. J. (1992). Optimality of sampling designs in item response theory models. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 274–288). Norwood NJ: Ablex Publishing.

Cook, R. D., & Nachtsheim, C. J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics, 22*, 315–324.

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.

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

Khan, M. K., & Yazdi, A. A. (1988). On D-optimal designs. *Journal of Statistical Planning and Inference, 18*, 83–91.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society, Series B, 21*, 271–319.

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

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

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*, 162–173.

Minkin, S. (1987). Optimal designs for binary data. *Journal of the American Statistical Association, 82*, 1098–1103.

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* (pp. 265–275). London: John Wiley & Sons.

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

Steinberg, D. M., & Hunter, W. G. (1984). Experimental design: Review and comment. *Technometrics, 26,* 71–130.

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

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55,* 461–475.

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 Report 87-2). Enschede: University of Twente, Department of Education.

van der Linden, W. J. (1988). *Optimizing incomplete sampling designs for item response model parameters* (Research Report 88-5). Enschede: University of Twente, Department of Education.

Wald, A. (1943). On the efficient design of statistical investigations. *Annals of Mathematical Statistics, 14,* 134–140.

Wingersky, M. S., & Lord (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wynn, H. P. (1970). The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics, 41,* 1655–1664.

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56,* 39–54.

Manuscript received 6/10/91
Final version received 12/16/91