Journal of Educational Statistics
Spring 1994, Vol. 19, No. 1, pp. 43–56

D-Optimal Sequential Sampling Designs for Item Response Theory Models

Martijn P.F. Berger
University of Twente, The Netherlands

Key words: *item response theory models, optimal design, sequential sampling*

The selection of optimal designs in IRT models encounters at least two problems. The first problem is that Fisher's information matrix is generally not independent of the values of the IRT parameters, and the second problem is that the design points are unknown parameters and have to be estimated together with the other parameters. In this study, these two problems are taken care of by a sequential design procedure. This procedure is a modification of a D-optimality procedure proposed by Wynn (1970). The results show that this algorithm leads to consistent estimates and that errors in selecting the abilities generally do not affect optimality very much.

Modifications or variations in designs have often been used to increase the information in the sample data for the estimation of parameters in a model. Such designs may lead to a reduction of the sample size needed to estimate the parameters efficiently. In educational measurement, this will often result in a reduction of the cost of testing.

The optimal designing of tests and samples in item response theory (IRT) models has been studied by various researchers. In all these studies, a certain function of Fisher's information matrix is maximized. Wingersky and Lord (1984), Lord and Wingersky (1985), Thissen and Wainer (1982), and de Gruijter (1985, 1988) studied the relative efficiencies of tests and models by means of the asymptotic variances of the estimated item parameters. Stocking (1990) selected optimum ability levels for item calibration by considering information on each of the item parameters separately. W. J. van der Linden (1988) maximized information on an item parameter by linear programming, and Vale (1986) used a relative efficiency measure to link item parameters. In addition to the above applied functions, Berger (1991) studied the efficiency of sampling designs and applied a determinant criterion function that not only considers the information on the parameters separately but also takes into account the joint information (covariation) among the parameters.

There are, however, some problems connected with the optimal designing of samples. One of the main problems in finding optimal designs for non-

linear IRT models is that Fisher's information matrix generally depends on the values of the unknown parameters—that is, one will first have to know the unknown parameters before one can actually select an optimal design to estimate these parameters. Several procedures have been proposed to handle this problem. See Berger and van der Linden (1992) for a brief review of these procedures.

One of the ways to deal with this problem is to estimate the parameters sequentially. Sequential procedures have been studied for a variety of models by Ford and Silvey (1980); Ford, Titterington, and Wu (1985); Wu (1985a); and McLeish and Tosh (1990), among others. Wu (1985b), for example, suggested a sequential design and parameter estimation procedure that has connections with the traditional stochastic approximation procedure (Robbins & Monro, 1951). This procedure is consistent and asymptotically normal under rather restrictive conditions. Berger (1992a, 1992b) describes some two-stage and sequential designs for IRT models. These procedures also make it possible to sequentially update initial estimates. In the following two practical settings, such a sequential design procedure may be very effective.

Builders of item banks not only have to conduct costly sessions to calibrate items for inclusion in an item bank but also face the problem that their item banks become exhausted—that is, already calibrated items become out of date or overexposed and have to be replaced by others. Usually, the calibration of these items takes place by administering them to a large fixed sample of examinees. Some parts of the thus-obtained sample data, however, may contain very little information on the item parameters, and the cost of such a large scale calibration of items may be reduced considerably by administering these items in separate stages. In the first stage, for example, a small sample of examinees may be used to obtain initial estimates of the item parameters, and in subsequent stages the items may be administered to a selected sample of examinees that will be more informative with respect to the estimated item parameters. In this way, efficient estimates of the item parameters may be obtained with a much smaller number of examinees.

A sequentially selected sample may also produce more efficient estimates of item parameters in computerized testing programs. These programs rely heavily on the availability of efficiently estimated item parameters. This may be obtained by choosing items for administration in a pretest mode on the basis of some crude knowledge of examinee ability. Examinees with a high contribution to the precision of the item parameter estimates may be selected sequentially in a pretest mode to calibrate the items.

In this article, a sequential design procedure is proposed that will not only sequentially design an optimal sample but is also able to sequentially estimate the item parameters. Such a procedure will be useful in large scale calibration studies, but it may also be important to test constructors and

builders of item banks and in computerized testing programs. The results of this article show that the procedure leads to consistent estimates and that optimal designs can be found relatively easily. First, however, a brief description of sampling designs for IRT models will be given.

Sampling Designs for IRT Models

Item response theory models assume that the probability of a response U_{ij} of an examinee j to an item i is a function of the latent abilities $\theta_j \in \mathbb{R}$. This probability will be denoted by $P_i(\theta_j)$ and can be represented by a parametric function of two item parameters:

$$P_i(\theta_j) = F\{a_i(\theta_j - b_i)\}, \tag{1}$$

where $F\{z\}$ may be a logistic or a normal ogive function. Item i is represented 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. For a whole test of n items, the n pairs $\{a_i, b_i\}$ together form the pair of vectors $\{\mathbf{a}, \mathbf{b}\}$.

A sampling design for IRT models is denoted by the pair of vectors $\{\boldsymbol{\theta}, \mathbf{W}\}$, where $\boldsymbol{\theta}$ consists of distinct abilities grouped as $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]'$ and the vector $\mathbf{W}$ consists of the corresponding weights $\mathbf{W} = [w_1, w_2, \ldots, w_c]'$. Note that the sample size is equal to $N = \Sigma_j w_j$. The elements of $\boldsymbol{\theta}$ are often called design points. By means of the pair of vectors $\{\boldsymbol{\theta}, \mathbf{W}\}$, a whole set of different sampling designs can be distinguished. When, for example, the number of distinct abilities is equal to the sample size—that is, $c = N$ and $w_j = 1$ for all $j = 1, 2, \ldots, c$—then all examinees in the sample have different abilities. If, however, $1 \leq c \leq N$, and all but one of the c weights are zero, then the sample consists of examinees all having the same abilities. For those cases where the abilities cannot be selected exactly but can be drawn from relatively homogeneous ability categories, it will be assumed that the pair of vectors $\{\boldsymbol{\theta}, \mathbf{W}\}$ denotes a sampling design where the abilities are drawn by a stratified sample from a population of abilities. In this case, the c elements θ_j are considered to be the c group (strata) means.

The asymptotic efficiency of the maximum likelihood estimators of the two item parameters for the logistic function is related to Fisher's information matrix:

$$J(a_i, b_i \mid \boldsymbol{\theta}, \mathbf{W}) = \sum_{j=1}^{c} w_j\{P_i(\theta_j)[1 - P_i(\theta_j)]\}\mathbf{X}\mathbf{X}', \tag{2}$$

where

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

and the total information matrix for a whole test with n items is a super-diagonal matrix $J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})$, with main diagonal matrices given by (2).

Although several functions defined on the space of the information matrix have been proposed in optimal design literature, the D-optimality (Kiefer, 1959), or generalized variance criterion (Anderson, 1984), may be preferred because it has some nice properties (see Berger, 1991, 1992a). An expression for the D-optimality criterion is given by:

$$\text{Det}[J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})] = \prod_{i=1}^{n} \text{Det}[J(a_i, b_i \mid \boldsymbol{\theta}, \mathbf{W})], \tag{3}$$

where $\text{Det}[J(a_i, b_i \mid \boldsymbol{\theta}, \mathbf{W})]$ is the determinant of the 2×2 information matrix for the item parameters a_i and b_i.

In the next sections, the sequential generation of an optimal design and the sequential estimation of parameters for IRT models will be discussed.

Sequential Generation of Optimal Designs

One of the most straightforward procedures to find optimal designs is to compute an optimality criterion for all possible combinations of parameter values and weights. Apart from the fact that an optimal design for one combination of parameters may not be optimal for another (i.e., designs are locally optimal), such a procedure is often inefficient in terms of CPU time. This is why Berger (1992a, 1992b) suggested the following optimal design generation procedure, which is an extension of an algorithm proposed by Wynn (1970).

The procedure is based on the ability vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]'$, containing all possible (grouped) abilities and the vector of weights $\mathbf{W} = [w_1, w_2, \ldots, w_c]'$, with an a priori chosen maximum sample size N_m. The procedure starts with an initial set of abilities $\boldsymbol{\theta}^{(0)}$ and corresponding weights $\mathbf{W}^{(0)}$. Design points are added sequentially after each stage, and the maximization procedure stops when $\Sigma_j w_j > N_m$. Thus, the maximization of $\text{Det}[J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})]$ is only constrained by the sample size N_m.

In the $(k + 1)$th step, the value $\theta_j^{(k+1)}$ with weight $w_j^{(k+1)}$ is selected out of the c abilities in $\boldsymbol{\theta}$ that has the largest value for:

$$\prod_{i=1}^{n} \{\text{Det}[J(a_i, b_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{W}^{(k)}) + w_j^{(k+1)} P_i^{(k+1)} Q_i^{(k+1)} \mathbf{X}^{(k+1)} \mathbf{X}'^{(k+1)}]\} \tag{4}$$

where $J(a_i, b_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{W}^{(k)})$ is Fisher's information matrix based on the previous k steps. $\mathbf{X}'^{(k+1)} = [-(\theta_j^{(k+1)} - b_i), a_i]$ and $P_i^{(k+1)} = [1 - Q_i^{(k+1)}]$, where $P_i^{(k+1)}$ is the probability of obtaining a correct response for $\theta_j^{(k+1)}$. Such a maximization will approximately lead to optimality under the condition that $\{J(a_i, b_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{W}^{(k)})\}^{-1}$ is nonsingular for each step.

This procedure can be applied to obtain an optimal design for a fixed set of parameters $\{\mathbf{a}, \mathbf{b}\}$. As an example, consider the three tests in Table 1 with $n = 9$ items each. These three tests cover a wide range of possible combinations of the item parameter values encountered in practice. The probability

TABLE 1
Three typical tests

Test A			Test B			Test C		
Item	a_i	b_i	Item	a_i	b_i	Item	a_i	b_i
1	0.2	−1.0	1	0.5	−2.0	1	0.5	−1.5
2	0.2	0.0	2	1.0	−1.5	2	1.0	−1.5
3	0.2	1.0	3	1.5	−1.0	3	1.5	−1.5
4	1.0	−1.0	4	0.5	−0.5	4	0.5	0.0
5	1.0	0.0	5	1.0	0.0	5	1.0	0.0
6	1.0	1.0	6	1.5	0.5	6	1.5	0.0
7	1.6	−1.0	7	0.5	1.0	7	0.5	1.5
8	1.6	0.0	8	1.0	1.5	8	1.0	1.5
9	1.6	1.0	9	1.5	2.0	9	1.5	1.5

mass functions of the corresponding locally optimal sampling designs are presented in Figure 1. The logarithms of the values for the Det$[J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})]$ of each of the tests are also given in Figure 1. These values are computed for the sample size of $N_m = 5{,}000$. The results show that, in general, a more or less uniformly distributed sample of abilities would be very efficient, and this conforms with the conclusions drawn by Wingersky and Lord (1984) and Stocking (1990).

The above-described procedure of first generating an optimal design sequentially and then estimating the parameters is not only very flexible but also makes it possible to use one model for design generation and another model or procedure for parameter estimation. It can also be applied to those cases where the abilities are not exactly known and where the examinees can

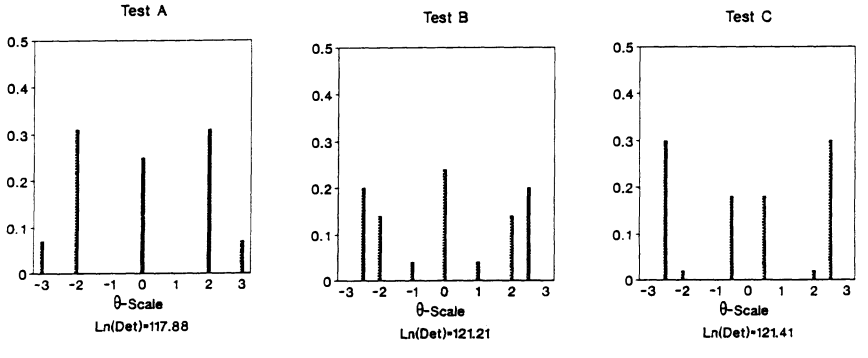

FIGURE 1. *Probability mass functions for the locally D-optimal sampling designs of three tests*

be grouped into categories with relatively homogeneous ability levels. In these cases, examinees can be sampled from the ability category that has the largest value for (4). From experience, it can be inferred that the optimal design is approached very well in these cases.

Another feature of this procedure is that its sequential nature makes it possible to estimate the parameters sequentially and use these updated estimates to refine the generation of the optimal design. In the next section, such a sequential design and estimation algorithm will be presented.

Finally, it should be mentioned that the flexibility of this procedure will also enable a researcher to use alternative forms having different weights in each step. If, for example, the weight $w_j^{(k+1)} = 1$, then the procedure may be considered to be fully sequential, and parameter estimates can be updated after each included observation. Although such a procedure has been applied in other areas (see Ford & Silvey, 1980), a fully sequential procedure will generally not be very practical in IRT modeling. When the weights $w_j^{(k+1)} > 1$, then a whole batch of $w_j^{(k+1)}$ design points is included in each step. This will not only speed up the iterations but may also improve the efficiency.

Sequential Estimation Procedure

In this section, the sequential design procedure will be combined with the sequential estimation of the parameters. The procedure starts with an initialization phase and stops when an a priori chosen maximum sample size N_m is reached.

Initialization Phase

-Set the maximum number of design points—that is, the maximum sample size—equal to N_m. The choice of N_m is usually based on a prior chosen minimum precision level m.

-Select a vector of all possible design points (abilities) $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]'$ with a fixed number of distinct ability levels or classes from which design points can be selected.

-Select an initial set of design points $\boldsymbol{\theta}^{(0)}$ with weights $\mathbf{W}^{(0)}$, such that $N^{(0)} = \Sigma_j w_j$. A suitable choice for such an initial set is a set of uniformly distributed design points—that is, with all the same weights w_j.

-Choose an initial set of (estimated) item parameters $\{\mathbf{a}^{(0)}, \mathbf{b}^{(0)}\}$.

-Obtain responses for the $N^{(0)}$ initial design points $\boldsymbol{\theta}^{(0)}$.

-Set $N^{(1)} = N^{(0)}$, $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$, and $\mathbf{W}^{(1)} = \mathbf{W}^{(0)}$.

k*th Iteration*

-Estimate the item parameters $\{\mathbf{a}, \mathbf{b}\}$ *bij* means of all the $N^{(k)}$ design points.

-Select the value $\theta_j^{(k+1)}$ with corresponding weight $w_j^{(k+1)}$ out of all possible design points in $\boldsymbol{\theta}$ that have the maximum value for (4). Note that, when

$w_j^{(k+1)}$ increases, the procedure will be speeded up, and that convergence may be reached much sooner.

-Add $w_j^{(k+1)}$ (approximate) design points $\theta_j^{(k+1)}$ to the previous design points in $\boldsymbol{\theta}^{(k)}$—that is, $N^{(k+1)} = N^{(k)} + w_j^{(k+1)}$.

-Obtain responses for the $w_j^{(k+1)}$ design points $\theta_j^{(k+1)}$.

-If $\Sigma_j w_j < N_m$, then continue the iteration; otherwise, stop.

It should be emphasized that both the choice of initialization variables and the selection of the weights $w_j^{(k+1)}$ in each iteration will influence the final outcome. But the differences are often very small, especially when the number of iterations is high. The choice of N_m may be based on the cost of obtaining responses and on the CPU time needed for the estimation of the parameters. An easy choice for the minimum precision level may be $m = \lambda_n^{-1}$, where λ_n is the smallest eigenvalue of the matrix $J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})^{-1}$, with a priori chosen variances of the estimators on the main diagonal. Because a uniformly distributed sampling design is often very efficient, an appropriate choice for the initial set $\{\boldsymbol{\theta}^{(0)}, \mathbf{W}^{(0)}\}$ may be a uniformly distributed sample. The initial sample size should be $N^{(0)} > N_a$, where N_a is the minimum sample size needed for the ML estimators to exist and to be identifiable and for $J(\hat{\mathbf{a}}, \hat{\mathbf{b}} \mid \hat{\boldsymbol{\theta}}^{(0)}, \hat{\mathbf{W}}^{(0)})^{-1}$ to be nonsingular. Finally, the number of design points added in each step will often depend on the type of application. If, for example, the number of iterations is limited to 10, then the weights will become $w_j^{(k+1)} = N_m/10$.

Consistency of Sequentially Estimated IRT Parameters

In general, there is no asymptotic theory available to support the use of likelihood confidence regions when data are sampled sequentially. There are, however, some results on the consistency of ML estimators available for a few cases. Wu (1985a) proposed a sequential procedure and demonstrated its consistency and asymptotic normality under restrictive conditions. Ford, Titterington, and Wu (1985) showed that the usual distributional results are also often valid for sequential sampling designs, and Wu (1985b) discussed some properties for binary data. The consistency of procedures of the kind presented in this article has been discussed by Tsay (1976) and Wu and Wynn (1978). Recently Chaudhuri and Mykland (1993) gave some asymptotic results for two-stage designs which validate inferences on the sequential estimator, ensure the optimality of the sequentially obtained design, and guarantee efficient parameter estimates.

If one has a sample of N observed responses and the ML estimates of $\{\mathbf{a}, \mathbf{b}\}$ are available with a known variance-covariance matrix $J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})^{-1}$, then the sampling distribution of the ML estimator $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\}$ will have nice properties asymptotically and will be strongly consistent—that is, $\lim_{N \to \infty} \{\hat{\mathbf{a}}, \hat{\mathbf{b}}\} = \{\mathbf{a}, \mathbf{b}\}$. If, however, $J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}, \mathbf{W})$ is unknown, and estimation of the parameters $\{\mathbf{a}, \mathbf{b}\}$

takes place sequentially, then updates of $J(\hat{\mathbf{a}}, \hat{\mathbf{b}} \mid \boldsymbol{\theta}, \mathbf{W})$ can be obtained in each step.

Suppose that $\hat{\lambda}_n$ is the smallest eigenvalue of $J(\hat{\mathbf{a}}, \hat{\mathbf{b}} \mid \boldsymbol{\theta}, \mathbf{W})$ and indicates the precision level for the estimates based on a certain number of design points. If m is the required minimum precision value, then the sequential procedure will stop when the sample size $N_m \in N$ is achieved, where

$$N_m = \inf\{N: \hat{\lambda}_n^{-1} > m; N > N_a\}. \tag{5}$$

N_a is the minimum sample size needed for $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\}$ to exist. Grambsch (1989) showed that under fairly general regularity conditions the sequential estimator $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\}$, which is based on N_m design points, is strongly consistent—that is:

$$\lim_{m \to \infty} \{\hat{\mathbf{a}}, \hat{\mathbf{b}}\} = \{\mathbf{a}, \mathbf{b}\} \qquad \text{(almost surely)}. \tag{6}$$

It may be inferred that a consequence of such a consistency is that

$$J(\hat{\mathbf{a}}, \hat{\mathbf{b}} \mid \boldsymbol{\theta}, \mathbf{W}) \to J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}^*, \mathbf{W}^*) \qquad \text{as } N \to \infty, \tag{7}$$

where $\{\boldsymbol{\theta}^*, \mathbf{W}^*\}$ is the optimal design for the estimation of $\{\mathbf{a}, \mathbf{b}\}$. These results indicate that one may expect the sequential estimation of the item parameters to lead to consistent results when the design points $\boldsymbol{\theta}$ are known. For those cases where $\boldsymbol{\theta}$ is unknown, but where the design points can be sampled from certain homogeneous classes with mean abilities $\boldsymbol{\theta}$, consistency may only be achieved when the sampling of design points is reasonably accurate.

When the design points $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_c)'$ are unknown and have to be estimated (approximated), then loss of information on the parameters a_i and b_i will be inevitable. Such a loss of information is caused by the fact that the optimal design points $\boldsymbol{\theta}^*$ cannot be located exactly. Analogous to the missing information principle of Orchard and Woodbury (1972), the following inequality for the diagonal elements of the information matrix will hold:

$$\text{Diag}\{J(\mathbf{a}, \mathbf{b} \mid \boldsymbol{\theta}^*, \mathbf{W}^*)\} \geq \text{Diag}\{J(\mathbf{a}, \mathbf{b} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}})\}, \tag{8}$$

where $\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{W}}\}$ is a design that is obtained from the sequential design procedure based on the (random) sampling of abilities (design points) from the ability categories $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_c)'$. Although such an inequality cannot be shown to hold for the determinant of the information matrix, one may expect the determinant of the information matrix based on estimated or approximated design points to be smaller than the corresponding determinant for known design points.

A Simulation Study

In this section, the results of some simulations will be given. The purpose of these simulations is to show that the proposed sequential procedure will

be consistent when the abilities (design points) are not exactly known but are approximated or selected randomly from distinct ability groups. Because it is often possible in achievement testing to group examinees into distinct but relatively homogeneous ability groups (strata), selection of these design points can take place by a stratified random sampling procedure. Although several tests with different combinations of item parameters were considered, only the results of the three tests from Table 1 will be presented, because the results were all very similar. The tests in Table 1 consist of combinations of parameter values ranging from $a_i = 0.2$ to $a_i = 1.6$ and $b_i = -2.0$ to $b_i = 2.0$.

Initialization phase. The maximum sample size was set equal to $N_m = 5{,}000$, and the initial sample consisted of all possible design points $\boldsymbol{\theta}^{(0)} = [-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3]'$, with weights all approximately equal to $w_j = N^{(0)}/c$, where $c = 13$. Three different forms were considered. The first form started with an initial sample size $N^{(0)} = 50$ and increased the sample after each iteration with $w_j^{(k+1)} = 50$ design points. The second form started with $N^{(0)} = 125$ and included $w_j^{(k+1)} = 125$ design points after each iteration. Finally, a third form was used with $N^{(0)} = 500$ and $w_j^{(k+1)} = 500$ to ensure a stable estimation procedure. The total number of iterations for these three forms was $N_m/50 = 100$, $N_m/125 = 40$, and $N_m/500 = 10$, respectively.

Data generation. After each iteration, examinees were randomly drawn from a uniform distribution of abilities with class mean values given by the θ_js in $\boldsymbol{\theta}^{(0)}$. Binary responses on item i for the examinees from ability level (class) θ_j were obtained by:

$$x_{ij} = \begin{cases} 1, \text{ if } P_i(\theta_j) > u_{ij} \\ 0, \text{ if } P_i(\theta_j) \leq u_{ij}, \end{cases}$$

where u_{ij} is randomly drawn from a uniform $U(0, 1)$ distribution and $P_i(\theta_j)$ is given by (1), with $F\{z\}$ being a logistic function.

Consistency measure. The most commonly used measure of performance of an estimator is the mean squared error (MSE). For tests consisting of n items which are characterized by the pair of vectors $\{\mathbf{a}, \mathbf{b}\}$, the mean squared errors are:

$$\text{MSE}_a = \sum_{i=1}^{n} \{E[(\hat{a}_i - a_i)^2]\}/n,$$

and

$$\text{MSE}_b = \sum_{i=1}^{n} \{E[(\hat{b}_i - b_i)^2]\}/n. \tag{9}$$

The ML estimates of the item parameters were placed on the scale of the true parameters by the linear transformation proposed by Stocking and

Lord (1983). The mean squared errors in (9) of the rescaled estimates are a function of both the variance of the estimators and their bias. These MSEs will eventually decrease to zero as more and more examinees are incorporated into the sample (i.e., as the number of iterations increases), if and only if both the variance of the estimators and their bias decreases as the sample size increases.

Results. In Figure 2, the results for both the parameters a_i and b_i are given for each of the three tests A, B, and C for the first 10 iterations. The MSE values are based on five replications. The $N^{(0)} = 50$ and $N^{(0)} = 125$ forms are less stable than the $N^{(0)} = 500$ form and generally have higher MSE values. Very inaccurate initial estimates in the $N^{(0)} = 50$ and $N^{(0)} = 125$ form continue to play an important role in subsequent iterations, because the initial set of examinees on which these estimates were based remains to be a relatively large part of the sample and because fewer examinees are added after each iteration. This leads to the unstable MSE pattern shown in Figure 2 for Test A.

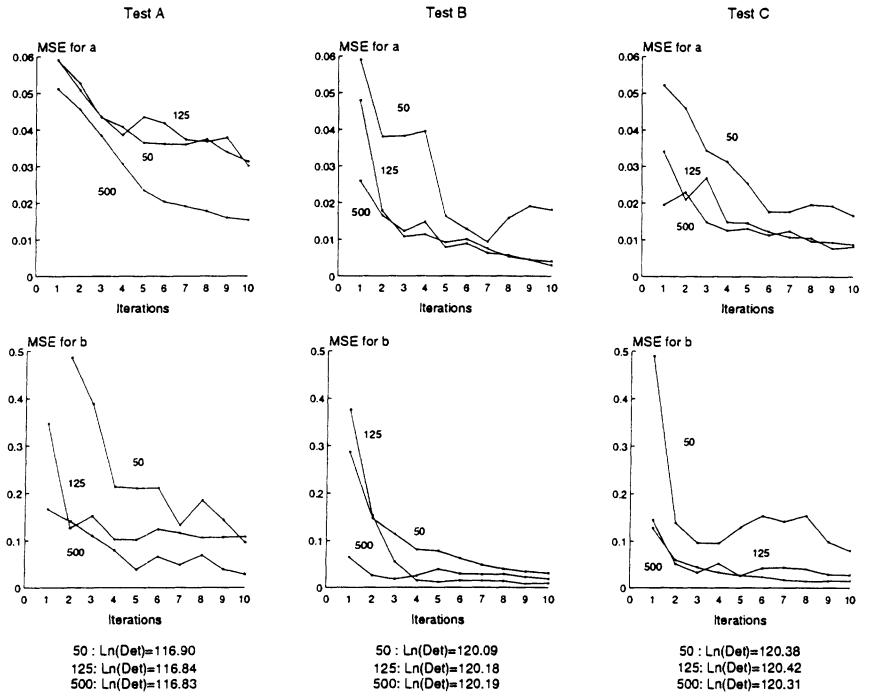

FIGURE 2. *Mean squared errors among the parameters and their sequential estimates for three tests*

In all cases, the MSE values decreased as the sample size (i.e., the number of iterations) increased. This means that the sequential estimation and design procedure are fairly consistent, though convergence seems rather slow in some cases.

An indication of the relative efficiency of the designs that were generated by this procedure can be obtained by relating the logarithms of the determinant of the information matrix $J(\mathbf{a}, \mathbf{b} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}})$ to the logarithms of the determinant obtained from the sequential design procedure with known item parameters and design points, which are given in Figure 1. The finally obtained $ln\,\{\mathrm{Det}[J(\mathbf{a}, \mathbf{b} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}})]\}$ values for the three forms are presented in Figure 2 and show that not much efficiency in estimating the item parameters is lost when the design points (abilities) are approximated. The $N^{(0)} = 50$ and the $N^{(0)} = 125$ forms seem to produce a somewhat higher criterion value than the $N^{(0)} = 500$ forms for tests A and C. Although these two forms start with higher MSE values than the $N^{(0)} = 500$ forms for tests A and C, the final result in terms of the determinant criterion is somewhat more optimal. The higher flexibility of these two forms may explain this effect. Note that these values were all computed for the same maximum sample size $N_m = 5{,}000$.

It must be noted that the use of other initialization variables and starting weights approximately leads to the same criterion values as shown in Figure 2, although minor differences in the actual optimal samples can be found. These differences occur because the procedure does not lead to one unique solution. In most applications, however, this will not be a problem because one is merely interested in finding an optimal design that enables efficient estimation of the item parameters.

Finally, it should be mentioned that the total amount of CPU time varied with the sample size but mainly consisted of the time needed to (re)estimate the parameters. Since the results were already very accurate after four or five iterations, the algorithm can be speeded up considerably by limiting the number of iterations to about five. An increase of the number of examinees to be added after each iteration will also make the algorithm faster.

Discussion and Conclusion

Item calibration is often a laborious and expensive process, and the use of more informative samples may decrease the cost of testing considerably. In this article, a procedure to locate D-optimal sampling designs for the efficient estimation of item parameters in IRT models is discussed. This procedure is very flexible and enables a researcher to increase the information on the item parameters in various ways.

First, the procedure may be applied fully sequentially—that is, the parameter estimates are updated after each single included design point. Although such a strategy may be useful, it is often very time consuming and generally not efficient for IRT models because the estimation of IRT

parameters is relatively expensive in terms of CPU time. Another approach is to reestimate the item parameters after the inclusion of a whole batch of design points. Such a procedure is often referred to as a multistage design procedure. A special case is the two-stage design procedure, which has only an initial and a final stage, where the design points are selected on the basis of the estimated item parameters in the first stage. From the results in this article, it may be concluded that a two-stage or a three-stage design procedure will generally lead to efficient results.

A comparison of the criterion values presented in Figure 2 with those presented in Figure 1 reveals that the criterion values do not differ much and that little efficiency is lost when the design points are approximated. Although the presented procedure generally results in an increase of efficiency, the actual increase, of course, will depend on the set of item parameters and the number of items in the test. Consider, for example, the optimal criterion values 117.88, 121.21, and 121.41 for the three tests A, B, and C in Figure 1. The corresponding values based on a sample of $N_m = 5{,}000$ examinees having $N(0, 1)$ distributed abilities are 113.65, 116.19, and 116.54, respectively. To obtain the optimal criterion values of the three locally optimal designs given in Figure 1, the normally distributed ability sample would need about 6,500 examinees. This means an increase of one fourth of the sample size.

The results of this article also show that the procedure will generally lead to consistent ML estimates of the item parameters. Although the results are limited to those cases where the same number of design points are added in each step, varying the number of included design points per step will generally not lead to different results. Including fewer design points in each step will only lead to a slower convergence rate.

Finally, it must be emphasized that this sequential procedure will not work very well for small samples. There is a sort of trade-off between the number of examinees included in each iteration and the total amount of CPU time, and we do not recommend including less than 100 examinees in each iteration.

References

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

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. (1992a). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*, *57*, 521–538.

Berger, M. P. F. (1992b). Generation of optimal designs for nonlinear models when the design points are incidental parameters. In Y. Dodge & J. Whittaker (Eds.), *Computational statistics* (Vol. 2, pp. 202–208). Heidelberg, Germany: Physica-Verlag.

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* (Vol. 1, pp. 274–288). Norwood, NJ: Ablex.

Chaudhuri, P., & Mykland, P. A. (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, *88*, 538–546.

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.

Ford, I., & Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, *67*, 381–388.

Ford, I., Titterington, D. M., & Wu, C. F. J. (1985). Inference and sequential design. *Biometrika*, *72*, 545–551.

Grambsch, P. (1989). Sequential maximum likelihood estimation with applications to logistic regression in case-control studies. *Journal of Statistical Planning and Inference*, *22*, 355–369.

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

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.

McLeish, D. L., & Tosh, D. (1990). Sequential designs in bioassay. *Biometrics*, *46*, 103–116.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In L. M. LeCam, J. Neyman, & E. L. Scott (Eds.), *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 697–715). Los Angeles: University of California Press.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *29*, 400–407.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

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

Tsay, J. Y. (1976). On the sequential construction of D-optimal designs. *Journal of the American Statistical Association*, *71*, 671–674.

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

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

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

Wu, C. F. J. (1985a). Asymptotic inference from sequential design in nonlinear situation. *Biometrika*, *72*, 553–558.

Wu, C. F. J. (1985b). Efficient sequential designs for binary data. *Journal of the American Statistical Association*, *392*, 974–984.

Wu, C. F., & Wynn, H. P. (1978). The convergence of general step-length algorithms for regular optimum design criteria. *The Annals of Statistics*, *6*, 1273–1285.

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

Author

MARTIJN P. F. BERGER is Professor, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. He specializes in optimal design, item response theory and computerized testing.

Received May 20, 1992
Revision received December 2, 1992
Accepted March 12, 1993