A General Approach to Algorithmic Design of Fixed-Form Tests, Adaptive Tests, and Testlets

Martijn P. F. Berger

University of Twente, The Netherlands

The selection of items from a calibrated item bank for fixed-form tests is an optimal test design problem; this problem has been handled in the literature by mathematical programming models. A similar problem, however, arises when items are selected for an adaptive test or for testlets. This paper focuses on the similarities of optimal design of fixed-form tests, adaptive tests, and testlets within the framework of the general theory of optimal designs. A sequential design procedure is proposed that uses these similarities. This procedure not only enables optimal design of fixed-form tests, adaptive tests, and testlets, but is also very flexible. The procedure is easy to apply, and consistent estimates for the trait level distribution are obtained. *Index terms: adaptive tests, consistency, efficiency, optimal test design, sequential procedure, test design, testlets.*

In education and psychology, there is a growing trend in the use of item banks. Achievement tests are composed of items selected from large banks of items having certain specifications. The selection of items from an item bank is facilitated by computerized systems, which are especially useful in large-scale assessment programs.

At least three different types of tests can be distinguished. The most traditional is the fixed-form conventional test that consists of a fixed set of items. These items may be drawn from an item bank with a priori calibrated items. A fixed-form test is neither equally valid nor equally reliable over the full trait range, which has stimulated the development of tests that are based on some type of "adaptation" of the test item to the trait level of the examinee taking the test. The work of Lord (1971, 1980) and Weiss (1976, 1978), among others, laid the groundwork for the implementation of adaptive testing. However, several practical problems with existing adaptive procedures were not solved. Because in an adaptive test every examinee potentially takes a different test, the context of an item will differ across examinees. Moreover, the effect of item ordering and the lack of robustness against different item orderings also may play a role in adaptive testing.

These problems with adaptive tests led Wainer & Kiely (1987) and Wainer & Lewis (1990) to propose the use of testlets. Instead of administering single items in each step as is done in adaptive tests, they proposed administering multi-item testlets with a fixed number of predetermined paths that an examinee may follow. Testlets combine the advantages of adaptive testing with the advantages of fixed-form tests. Item response theory (IRT) models are especially useful for the analysis of the test responses. These models have not only been very successful in fixed-form conventional tests but also in adaptive tests and testlets.

The purpose of administering achievement tests is to obtain an efficient estimate of the trait levels of a sample of examinees. The trait estimates obtained from different test types, however, may differ in efficiency. The purpose of the present study was to stress the similarities in designing fixed-form tests, adaptive tests, and testlets using optimal design theory. Reviews of optimal design theory are given by Silvey (1980) and Ford, Kitsos, & Titterington (1989).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 2, June 1994, pp. 141–153
© Copyright 1994 Applied Psychological Measurement Inc.
0146-6216/94/0146-6216/94/020141-13$1.90

A sequential design procedure that not only generates optimal test designs for fixed-form tests and adaptive tests, but also enables a researcher to find an optimal set of items for inclusion in testlets, is presented. This procedure originated from optimal design research and may serve as a fast and simple alternative to the more complicated mathematical programming algorithms that have been applied primarily to construct fixed-form tests. The sequential procedure takes into account the shape of the latent trait distribution of the sample of examinees.

Designing Optimal Tests

A test design is mainly characterized by the pattern of the examinee-item interaction. The selection of items and their administration is not only based on their capability of measuring individual differences, but also is guided by the substantive properties of the items. Several procedures have been proposed to construct a test, and most of these procedures start with some prior specified amount of information for each trait level. The actual construction of the test is then a matter of finding items that in some way match the specified information level. The most commonly used procedures are those based on mathematical programming, and these procedures have been applied in computerized test construction. See, for example, Adema (1990) and Boekkooi-Timminga (1989) for reviews of alternative mathematical programming models in test construction.

Notation and Assumptions

Suppose that an optimal test design for a sample of N examinees ($k=1, \ldots, N$) and n items ($i=1, \ldots, n$) is to be constructed. Let the response matrix be given by $\mathbf{u}=\{u_{ik}\}$. Assume that these responses are dichotomous and that the probability of obtaining these responses is given by the function $P(\theta_j;\xi_i)$, where θ_j is a parameter from a vector $\theta'=(\theta_1,\theta_2,\theta_3, \ldots, \theta_c)$, representing c distinct values of the trait (θ) scale, and $\theta_j \in \mathbb{R}$, where $\mathbb{R}$ is a set of real numbers. A vector of weights $\mathbf{W}'=(w_1,w_2,w_3, \ldots, w_c)$ corresponds with these values. These weights can be defined in different ways. Here it is assumed that the weights in $\mathbf{W}$ characterize the shape of the discrete θ distribution of the sample of N examinees. This means that $w_j \geq 0$ for all $j=1, \ldots, c$, $\sum_j w_j = N$, and $1 \leq c \leq N$. In other words, the pair of vectors $\{\theta,\mathbf{W}\}$ represents this discrete frequency distribution for θ, and the choice of weights determines the shape of this distribution.

For example, a uniform θ distribution with c categories in θ will have equal weights $w_j = c/N$ for each of the categories. On the other hand, if all $w_j = 0$ except one, then all examinees will have the same θ level. The same design will arise for $c=1$. Another possibility is that $c=N$ and that all $w_j=1$. Then all examinees in the sample will have different θ levels. It should be emphasized that because $\sum_j w_j = N$, all weights in $\mathbf{W}$ can be divided by N without loss of information. This means that N does not affect the optimality of a test design.

It also is assumed that a vector of parameters $\xi_i' = \left(\xi_{1,i}, \xi_{2,i}, \xi_{3,i}, \ldots, \xi_{P,i}\right)$ characterizes item i, where $\xi_i \in \mathbb{R}^P$, and $\mathbb{R}^p$ is a p-dimensional set of real numbers. The mean and variance of the parametric family are $\mathrm{E}\left\{\mu_{ij} \middle| \xi_i\right\} = P\left(\theta_j;\xi_i\right)$and $\mathrm{Var}\left\{\mu_{ij} \middle| \xi_i\right\} = P(\theta_j;\xi_i)\left[1-P(\theta_j;\xi_i)\right]$, respectively. An example from this family for $p=2$ is the two-parameter logistic model (2PLM)

$$P(\theta_j;\xi_i) = \left\{1+|\exp|\left[-\xi_{2i}(\theta_j - \xi_{1i})\right]\right\}^{-1}. \tag{1}$$

In computerized test construction the items first must be calibrated and placed in a structured item bank. It generally is assumed that a set of parameter vectors, $\xi' = \left(\xi_1', \xi_2', \xi_3', \ldots, \xi_n'\right)$, describes an n-item test, and that these parameters can be replaced by their estimates without loss of information. The purpose of test construction is to select a set of n items from the item bank by means of their estimated characteristics for efficient estimation of the parameter vector $\theta' = (\theta_1,\theta_2,\theta_3, \ldots, \theta_c)$. Under the assumption of local independence, the likelihood associated with the set of responses $\mathbf{u}$ and the vector θ is

$$L(\mathbf{u};\theta;\xi) = \prod_{j=1}^{c}\prod_{i=1}^{n} P_i(\theta_j)^{w_j p_{ij}} \left[1-P_i(\theta_j)\right]^{w_j(1-p_{ij})}, \tag{2}$$

where p_{ij} is the proportion of correct responses on item i in category θ_j. The Fisher information function connected with the parameter θ_j is defined by

$$I_u(\theta_j) \equiv \mathrm{E}_{\theta_j}\left\{\frac{\partial}{\partial\theta_j}\ln[L(\mathbf{u};\theta;\xi)]\right\}^2 . \tag{3}$$

The information $I_u(\theta_j)$ on the parameters in θ can be gathered in the vector $\mathbb{I}_u(\theta|\xi)$, $[\text{i.e., } \mathbb{I}_u(\theta|\xi)' = \{I_u(\theta_1), I_u(\theta_2), \mathrm{I}_u(\theta_3), \ldots, I_u(\theta_c)\}]$. Optimal test design uses Fisher's information and selects an optimal set of n test items with parameters ξ from the item bank in such a way that θ is estimated as efficiently as possible (i.e., the estimators of θ have the smallest variance). Generally, two problems arise.

First, the optimal design can only be determined exactly when all parameters in the IRT model are known. Good approximations of an optimal design can only be obtained when parameter values can be replaced by their estimates without loss of information. The second problem is that Fisher's information is a function of the θ parameters themselves, and an optimal test design thus depends on the values of the θ parameters. Different values for the θ parameters can lead to different optimal test designs. Thus, optimality can only be obtained locally (i.e., for a given set of θ parameters). This is why the term *local optimality* has been used in the literature on optimal designs. A definition for a locally optimal test design for a pair of vectors $\{\theta, \mathbb{W}\}$ is provided below. This definition is based on the function $\Phi(\cdot)$ of the vectors $\mathbb{I}_u(\theta|\xi)$ and $\mathbb{W}$.

Definition of Locally Optimal Test Design

A test design with a set of parameter vectors for n items, $\xi^{*\prime} = (\xi_1^*, \xi_2^*, \xi_3^*, \ldots, \xi_n^*)$, where $\xi_i^* \in \mathbb{R}^p$ and $\mathbb{R}^p$ is a p-dimensional set of real numbers, is locally Φ-optimal if

$$\Phi[\mathbb{I}_u(\theta|\xi^*), \mathbb{W}] \geq \Phi[\mathbb{I}_u(\theta|\xi), \mathbb{W}] \tag{4}$$

for a given pair of vectors $\{\theta, \mathbb{W}\}$, where $\theta \in \mathbb{R}^c$ with weights $\mathbb{W}$ and any $\xi' = (\xi_1, \xi_2, \xi_3, \ldots, \xi_n)$, where $\xi_i \in \mathbb{R}^p$.

In this definition, $\Phi(\cdot)$ represents an entire class of real-valued functions having certain properties; see Silvey (1980) for a review of these properties. A member of the family $\Phi(\cdot)$ is the product function

$$\Phi\{\mathbb{I}_u(\theta|\xi), \mathbb{W}\} = \prod_{j=1}^{c} I_u(\theta_j)^{w_j} , \tag{5}$$

or its equivalent, the sum of the log function

$$\Phi\{\mathbb{I}_u(\theta|\xi), \mathbb{W}\} = \sum_{j=1}^{c} w_j \log I_u(\theta_j) \cdot \tag{6}$$

The function in Equation 5 is equal to the D-optimality criterion used by Berger (1991). This can be verified easily by assuming that the vector $\mathbb{I}_u(\theta|\xi)$ is the main diagonal of a $c \times c$ diagonal matrix. The determinant or D-optimality criterion is then the (weighted) product of the diagonal elements. Of course, alternative criteria belonging to the family $\Phi(\cdot)$, such as the E-optimality and the A-optimality criteria (see Atkinson, 1982), or the MAXIMIN-optimality criterion (Van der Linden & Boekkooi-Timminga, 1989), also can be applied. A review of these alternative criteria is given by Berger & van der Linden (1992) and Berger & Veerkamp (in press).

As mentioned above, optimal test designs cannot be found easily. First, the sample of examinees usually has an unknown θ distribution $\{\theta, \mathbb{W}\}$. An optimal design for one θ distribution $\{\theta, \mathbb{W}\}$ may not be optimal for another and errors in specifying $\{\theta, \mathbb{W}\}$ may have a large effect on the optimality. Although the application of adaptive tests is done to circumvent this problem, errors in estimating each of the parameters in θ may continue to play a large role in adaptive testing.

Another problem is that, given an item bank with a very large number of calibrated items, the optimal set of n items with parameters ξ^* may be impossible to locate exactly. Even if additional constraints are imposed, the total number of combinations to consider may be extremely high. Because different combinations of items may lead to the same optimal criterion value, mathematical programming procedures can be used to locate such approximate optimal test designs (e.g., Adema, 1990; Boekkooi-Timminga, 1989; Theunissen, 1986; Van der Linden & Boekkooi-Timminga, 1989).

The effect of using estimates instead of parameter values, however, may lead to certain problems. A study by Hambleton, Jones, & Rogers (1993), for example, indicated that the test information function (TIF) could be overestimated even when large samples were used. This effect may be explained by capitalization on chance due to positive errors in the item parameter estimates, and may lead to underestimation of the variance of the θ estimates. However, no information is yet available about the seriousness of this effect in practical situations.

Locally Optimal Adaptive Test Design

The most frequently applied solution to the problem of not knowing the θ distribution of the sample exactly is to use an adaptive testing procedure. The problem of finding an optimal set of test items for the efficient estimation of a θ distribution in adaptive testing is similar to the problem of finding a fixed-form optimal test design. The only difference is that in adaptive testing the selection of items with maximum information is based on $\hat{\theta}$s obtained from responses to previously administered items. The definition given above for fixed-form tests can be changed to define a locally optimal adaptive test design by replacing $\mathbf{I}_u(\theta|\xi)$ with $I_u\left[\hat{\theta}^{(n)}\middle|\xi^{(n)}\right]$, where the $\hat{\theta}^{(n)}$s are the estimated θs after successive administration of n items with $\xi^{(n)}$ characteristics. Of course, $\hat{\theta}^{(n)}$ usually characterizes the θ distribution of a single examinee (i.e., $c = 1$, and θ and $W = 1$ are scalars).

Locally Optimal Testlet Design

Wainer & Kiely (1987) described the testlet as "a group of items related to a content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). Testlets are actually small tests, and they can be combined to form one long test. This makes it possible to design a test in such a way that examinees may be allowed to take only some of the items in each testlet, and may even be allowed to take only some of the testlets within the test. Wainer & Kiely (1987) reviewed alternative testlet forms. The optimal design of testlets can be done in two different ways—within-testlet optimality can be distinguished from between-testlet optimality.

A test may, for example, consist of six different testlets. If these testlets are ordered from easy to difficult, and each examinee begins with the easiest testlet and stops whenever a testlet is not completed successfully, then the resulting design reflects a pattern in which the most able examinees take more testlets than the less able examinees. The optimal selection of testlets for inclusion in such a test can be done by assuming that the item parameter vector, ξ_i, contains the item parameters of all items grouped within each testlet, and that ξ includes all the item parameters of the total test. Because optimality of the testlet design is obtained by selecting between alternative testlets, this optimality can be referred to as between-testlet optimality.

Within-testlet optimality is defined for the set of items within a testlet, and can be obtained by selecting a set of calibrated items from an item bank that has a maximum value for the function $\Phi(\cdot)$. In this case, ξ_i contains the item parameters of each item within the testlet, and ξ contains all the item parameters in the testlet. The paths through a testlet may follow a hierarchical or a linear structure. Depending on the structure of the testlet, the pair of vectors $\{\theta, \mathbf{W}\}$ have a different size and different upper and lower bounds in each step.

A Sequential Test Design Procedure

Several procedures to generate optimal designs have been proposed in the literature. Berger (1992a, 1992b, 1994) described a sequential sampling design procedure for the optimal selection of examinees for item calibration. The present procedure is a modification of this method. Instead of a sequential selection of examinees to obtain item parameter estimates, this procedure sequentially selects items from an item bank for the optimal estimation of θ.

The design procedure consists of sequential selection of (sets of) items out of the available items in the item bank such that a function from the class $\Phi(\cdot)$ has a maximum value in each step. The sequential nature of the procedure enables the researcher to update the θ parameter estimates in each step of the procedure. A main feature of the procedure is that it is very flexible and takes into account the actual shape of the θ distribution. It can be applied to generate optimal designs for fixed-form tests, adaptive tests, or testlets, and it can be applied to different IRT models. Finally, the sequential characteristics of the procedure ensure that the procedure leads to consistent estimators for the θs.

The Initialization Phase: Step (0)

An item bank with calibrated items is specified, and the maximum number of items n_m for the test is selected.

The initial frequency distribution $\{\theta^{(0)}, \mathrm{W}^{(0)}\}$ is specified. For an adaptive test or testlet, the distribution $\{\theta^{(0)}, \mathrm{W}^{(0)}\}$ will have restrictions on the range of the θ_j values.

An initial item with parameters $\xi_i^{(0)}$ is selected.

In an adaptive test, separate responses must be obtained at each step.

The (*k*+1)th Step

A set of items is selected using the parameter vector $\xi_i^{(k+1)}$ out of all possible items in the item bank that has a maximum value for

$$\Phi\left\{\left[I_u\left(\theta^{(k)}\middle|\xi^{(k)}\right)+I_u\left(\theta^{(k)}\middle|\xi_i^{(k)+1}\right)\right], \mathrm{W}^{(k)}\right\}. \qquad (7)$$

$\xi^{(k)}$ contains parameters from all previous selected items, and $\theta^{(k)}$ and $\mathrm{W}^{(k)}$ describe the θ distribution for the previous k steps.

In adaptive testing, the θs are re-estimated after each step; therefore, $\theta^{(k)}$ in Equation 7 is replaced by its estimate and will generally not be the same in each step because of the stepwise updating of θ. On the other hand, if the procedure is used for the generation of items for a hierarchical testlet, the range of $\theta^{(k)}$ will be different in each step. This is illustrated below. The procedure allows a single item to be included in each step or for an entire set of items to be included in each step. This is referred to as a multi-stage testing procedure; when the number of included items is $n_m/2$, the sequential procedure reduces to a two-stage testing procedure.

If the number of items is less than n_m, the iteration continues, otherwise the procedure stops.

Consistency of Sequential Estimators

The sequential design and estimation procedure includes item responses to sequentially update the θ parameter estimates. The implementation of this procedure in an adaptive testing environment may not only include estimation of the θs in each iteration, but may also include updating estimates of the parameters characterizing the selected items. The primary problem, however, is whether these obtained estimates have the same properties as the usual maximum likelihood estimates based on a fixed-form test design.

The consistency of sequential estimation procedures has been studied by many researchers. Wu (1985) and Ford, Titterington, & Wu (1985) showed that maximum likelihood (ML) estimators are consistent in some special cases. The sequential design procedure proposed here is comparable to a procedure proposed by Wynn (1970), and the consistency of Wynn's procedure has been proven by Wu & Wynn (1978) and Tsay (1976).

Suppose that a set of responses **u** is available for an n-item test, and suppose also that the ML estimates for θ have variances $\mathbb{I}_u(\theta|\xi)^{-1}$. Then standard large sampling theory indicates that the ML estimator of θ is strongly consistent, that is,

$$P(\lim\hat{\theta} = \theta) = 1, \quad as\ n \to \infty \cdot \tag{8}$$

Grambsch (1989) showed that under fairly general regularity conditions, the sequential estimator $\hat{\theta}^{(n)}$ based on n observations is strongly consistent; that is, $\lim\hat{\theta}^{(n)} = \theta$. More recently, Chaudhuri & Mykland (1993) provided more asymptotic results that validate statistical inference based on the sequential estimator $\hat{\theta}^{(n)}$. From this result it may be inferred that

$$I_u\left(\hat{\theta}^{(n)}\middle|\xi\right) \to I_u(\theta|\xi^*), \quad as\ n \to \infty \cdot \tag{9}$$

These results are valid only for long tests, and the question of how long the test should be for a valid application of sequential test designs has not yet been answered. This means that in adaptive testing, for example, inferences based on large sampling theory should be applied with caution. Particularly, the interpretation of information as the inverse variance of ML estimators in sequential design procedures is not generally justified. Thissen & Mislevy (1990), however, noted that experience with real and simulated data suggests that sequential design procedures do not substantially affect ML estimates or their precision. Sequentially obtained ML estimates seem to provide good final estimates, and the inverse of the value of the information function seems to be an adequate estimate of the variation in sequential adaptive testing.

Example Applications

The sequential design procedure was used to illustrate how optimal designs for fixed-form tests, adaptive tests, and testlets can be obtained. The results are based on the D-optimality criterion because it has some useful properties; see Berger (1991, 1992b) for an explanation of these properties. It should be emphasized, however, that other criteria belonging to the class of criteria $\Phi(\cdot)$ also may be used and may have their own advantages in certain applications.

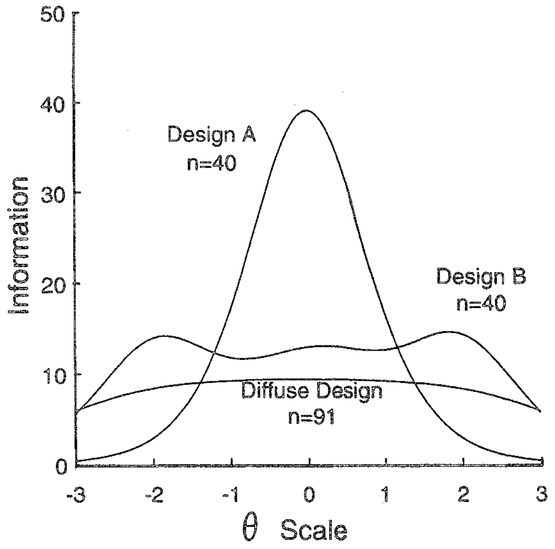
D-Optimal Fixed-Form Test Designs

For the sequential generation of a D-optimal fixed-form test design, an item bank was constructed with items having all possible combinations of item parameter values. Although the sequential procedure can be applied to any parametric IRT model, the items in the item bank were calibrated with a $p = 2$ parameter logistic model (2PLM); that is, with a parameter vector $\xi_i = (b_i, a_i)$, where b_i is the difficulty parameter that ranged from -3 to $+3$, and a_i is the discrimination parameter that ranged from $.5$ to 2.0, and continuous random parameter values were computer generated within these ranges. Tests with $n = 40$ items were generated by the sequential procedure.

The discrete θ distribution $\{\theta, \mathbb{W}\}$ for which the optimal test was selected was characterized by $c = 25$ distinct equally spaced classes $\theta = (-3.0, -2.75, -2.5, \ldots, 2.5, 2.75, 3.0)$. The corresponding weights $\mathbb{W}$ could be divided by the sample size N without loss of information, because an optimal design does not depend on the actual N.

The TIFs for two locally D-optimal test designs, Designs A and B, are shown in Figure 1. Design A was generated for a sample with a truncated θ distribution $\{\theta, W\}$, where all weights w_j were 0, except the equal weights connected with the θ values $-.25, 0.0$, and $.25$, respectively. Design B was a locally D-optimal test design for a sample with uniformly distributed θ values; that is, θ ranged from -3 to $+3$ and all weights $w_j = 1$. To illustrate that a substantial reduction of the number of items can be obtained, the TIF for a "diffuse" test consisting of $n = 91$ items drawn uniformly from all available combinations of b_i and a_i also is shown in Figure 1. A reduction of approximately 56% of the number of items was obtained when Design B was used instead of the diffuse test design. Over the entire range of θs, Design B displayed more information than the diffuse design.

Figure 1
TIFs for Two Locally D-Optimal Test Designs



The probability mass functions of three locally D-optimal test designs for a uniform sample of θs are shown in Figure 2. These probability mass functions represent the proportion of items in the test with a particular combination of item parameter values. For example, for $a_i = 1.0$, the locally D-optimal test design consisted of a test with approximately 40% of the items having $b_i = -1.0$ and 40% with b_i=1.0. The remaining 20% of the items in the optimal test had other values for b_i. These percentages remained the same for different test lengths. Depending on the size of a_i it can be seen that an optimal design is mainly characterized by a bimodal selection of items from the scale of b_i. Increasing values for a_i led to an increase of the difference between the two modes of the mass function.

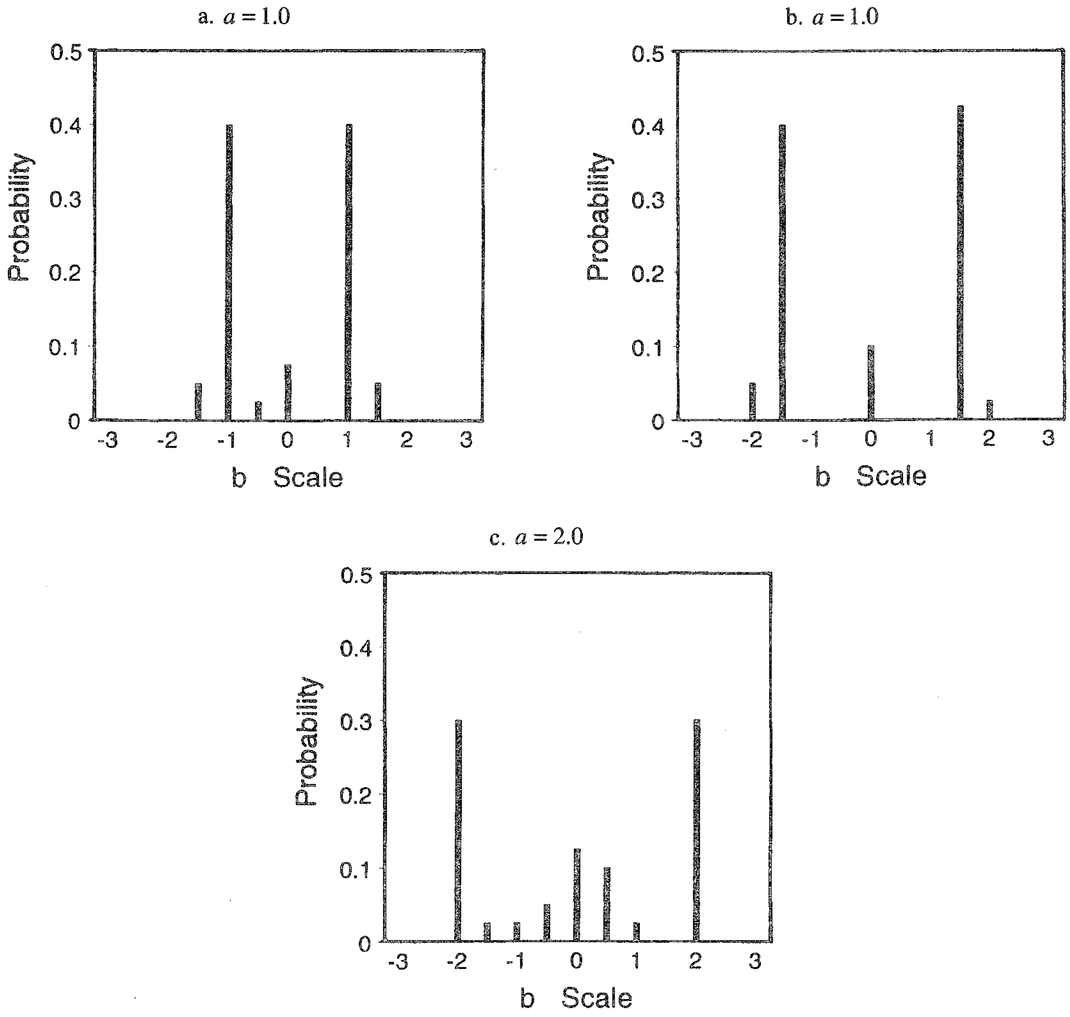
The corresponding TIFs are presented in Figure 3. The shape of these functions varied for different values of a_i. This is a result of the difference between the two modes of the mass functions. A similar result was obtained by Baker, Cohen, & Barmish (1988) for test assembly by means of linear programming. They also found a bimodal clustering of items with respect to b. Thus, Figures 2 and 3 show that the optimality of a test design for the 2PLM is primarily characterized by a bimodal selection of items from the b_i scale of the item bank, and that the TIFs become more bimodally shaped as the selected items have higher discriminatory power.

D-Optimal Adaptive Test Designs

The application of the sequential design procedure to adaptive testing is straightforward. In adaptive testing, a single examinee is considered—θ reduces to a scalar and $\mathbf{W} = 1$. Suppose that θ_x is the trait level of a single examinee x. Then, in adaptive testing, items are selected that have maximum information at a point estimate of θ_x. This estimate is based on the examinee's preceding responses. The success of an adaptive testing approach depends on the availability of accurate initial and sequential estimates of θ_x. In those cases in which the estimates of θ_x have large variances, the adaptive testing procedure may become unwieldy and even inconsistent. In other words, the uncertainty of the estimator of θ_x may disturb the procedure.

The sequential design procedure can take into account the uncertainty of the point estimate of θ_x by using a confidence interval for θ_x. Because the ML estimator $\hat{\theta}_x$ is asymptotically normally distributed with mean

Figure 2
Probability Mass Functions of Locally D-Optimal Test Designs for a Uniform Sample of θs

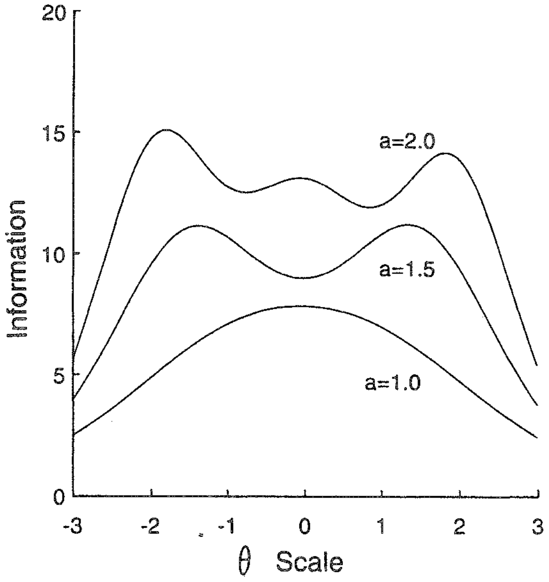


a. $a = 1.0$

b. $a = 1.0$

c. $a = 2.0$

Figure 3
TIFs of Locally D-Optimal Test Designs for a Uniform Sample of θs



θ_x and variance $I_u(\theta_x)^{-1}$, the limits of the approximate $(1-\alpha)\%$ confidence interval for θ_x can be formulated as

$$\theta_1 = \hat{\theta}_x - z_{(1-\alpha/2)} I_u(\theta_x)^{-1} \tag{10}$$

and

$$\theta_c = \hat{\theta}_x + z_{(1-\alpha/2)} I_u(\theta_x)^{-1} , \tag{11}$$

where $z_{(1-\alpha/2)}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. The D-optimality criterion now can be formulated using the pair of vectors $\{\theta, W\}$, where θ consists of all θ_j values within the confidence interval with a lower bound θ_1 and an upper bound θ_c. An alternative item selection criterion for adaptive testing that also is based on a confidence interval but uses different weights W was proposed by Veerkamp & Berger (1993).

Thus, errors in estimating θ_x values may lead to instability of the adaptive testing procedure, especially at the beginning of the item administration. The use of a confidence interval-based criterion, however, will make the procedure more robust against such estimation errors, and the problem of capitalization on chance caused by using estimates for the parameter values (Hambleton et al., 1993) also will be expected to diminish with such a criterion.

D-Optimal Testlet Designs

Although different kinds of testlets are possible (Wainer & Kiely, 1987; Wainer & Lewis, 1990), the procedure—designed to obtain an optimal selection of items for a testlet (i.e., to obtain within-testlet optimality)—was applied to a hierarchically structured testlet. Figure 4 shows three optimally selected sets of items for a hierarchical structured testlet. The items in Figure 4 are ordered according to their presentation order and their *b* levels. These items were selected from an item bank using the sequential procedure. The items in the item bank were calibrated using the 2PLM. The items had all possible combinations of item parameter values ranging from $b_i = -3.0$ to $b_i = 3.0$, and $a_i = .5$ to $a_i = 2.0$. The testlet designs in Figures 4a–4c are locally D-optimal for a negatively skewed distribution, a uniform distribution, and a normal distribution of $\{\theta, \mathbb{W}\}$, respectively.

Because the testlet was hierarchically structured, the sequential selection of items was performed by truncating the range of the vectors $\{\theta, \mathbb{W}\}$ on which the D-optimality criterion was based in each step. For example, when the procedure started with θ values ranging from –3.0 to 3.0, and the first optimally selected item had $b_i = 1.75$, as shown in Figure 4a for the negatively skewed distribution, then the next items were selected by means of the D-optimality criterion based on the truncated vectors $\{\theta, \mathbb{W}\}$, with θ values ranging from –3.0 to 1.75, and θ values ranging from 1.75 to 3.0, respectively. The vector $\mathbb{W}$ was truncated accordingly.

Figure 4 shows that although the first item selected generally was located near the median of the distribution, the sequentially selected items in the trees spread out further as the presentation order increased. The influence of the shape of the θ distributions also is displayed in Figure 4. For example, the spread of items over the b_i scale was higher for the uniform distribution (Figure 4b) than for the normal distribution (Figure 4c). This shows that the optimal selection of items in a testlet depends on the shape of the θ distribution of the sample for which the testlet is constructed.

Note that although the items selected for these testlets are only displayed by their *b* values, their *a* values also play a role. This explains why for the uniform design in Figure 4b the order 2 *b*s were not symmetric about the order 1 point, and why order 3 *b*s were not symmetric about the order 2 *b*s.

Discussion and Conclusions

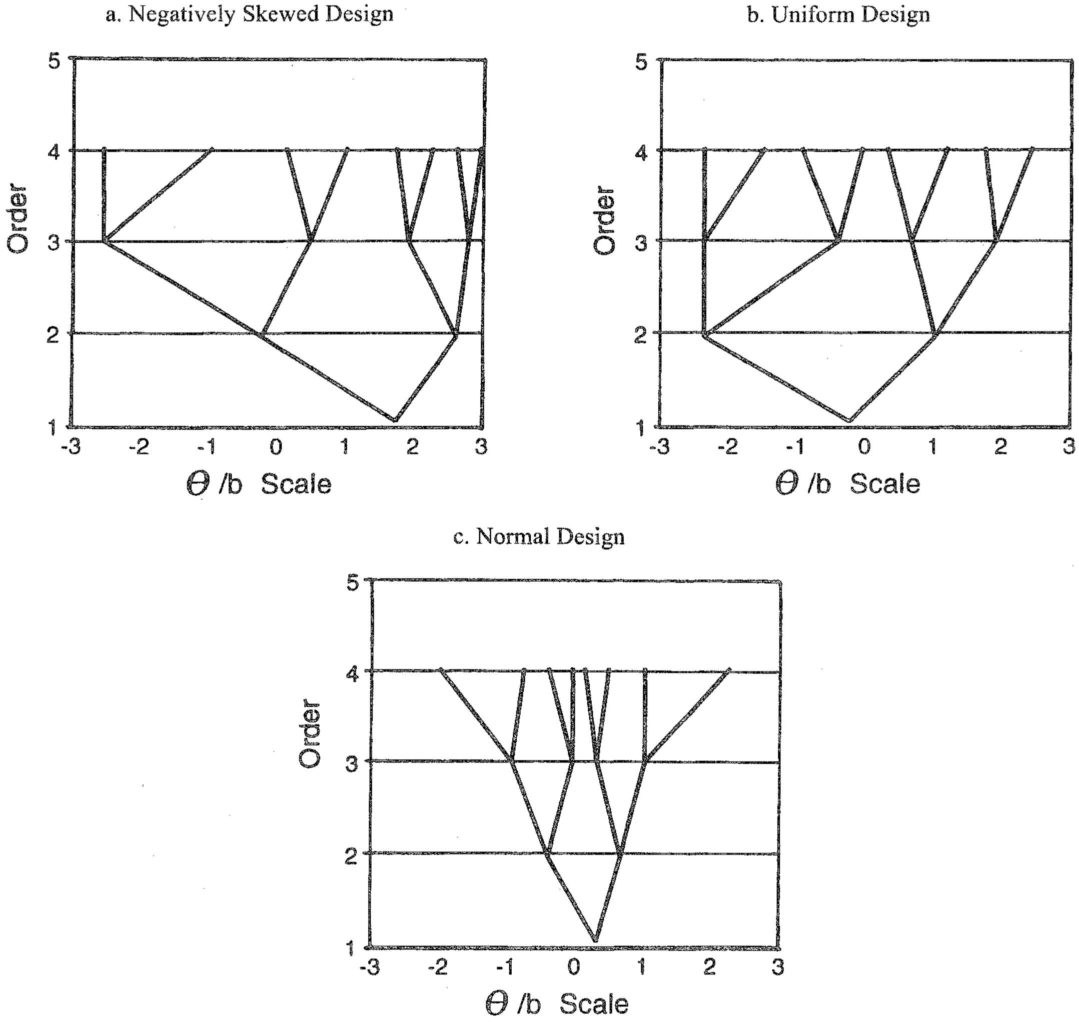
The results showed that the sequential test design procedure was very effective in obtaining optimal test designs, and its flexibility enables the test constructor to apply the procedure to different test types. This procedure is based on the pair of vectors $\{\theta, \mathbb{W}\}$ that characterize a discrete θ distribution of a group of examinees for whom the test is designed. The sequential nature of the procedure allows the parameter estimates to be updated in each step. This makes the procedure especially useful for adaptive testing. Because the objective function allows restriction of the θ range, the procedure also can be applied to optimally construct hierarchical testlets, by truncating the range of $\{\theta, \mathbb{W}\}$ in each step of the procedure. As such, this procedure may be considered to be more generally applicable than the mathematical programming models, which have been applied primarily to the construction of fixed-form tests.

Another important feature of this procedure is that it takes into account the shape of the θ distribution. The results showed that the shape of the θ distribution of the sample of examinees is important for the optimal selection of items, and that a bimodal selection of items on the *b* scale is often optimal for a uniform θ distribution. Baker et al. (1988) found that test assembly based on linear programming also displayed a bimodal pattern for the *b* parameter when a uniform target information function was used. Although no mathematical programming model has yet been formulated that directly takes into account the shape of the θ distribution of the group of examinees for whom the test is constructed, the target information function in linear programming may play a role similar to the role of the weights $\mathbb{W}$ in this general approach. A more detailed study of these similarities, however, is needed.

The sequential design procedure can be applied with any function belonging to the general class of functions $\Phi(\cdot)$. This class not only includes some of the criteria used in mathematical programming models, such as the MAXIMIN criterion (Van der Linden & Boekkooi-Timminga, 1989), but also includes criteria that have not yet been applied to test construction nor in mathematical programming models (Silvey, 1980). The generality of the sequential design procedure, however, has its limitations.

A major feature of mathematical programming models is that many practical constraints can be included relatively easily. A review of several requirements concerning test format and content formulated as linear

Figure 4
Three Locally D-Optimal Hierarchical Testlet Designs


a. Negatively Skewed Design
b. Uniform Design
c. Normal Design

constraints was given by Van der Linden & Boekkooi-Timminga (1989). Although it should be possible to incorporate these linear constraints into the present sequential design procedure, such extensions need to be studied in future research.

Finally, the procedure proposed here cannot be applied easily to compare the relative efficiencies of different test types. The success of any optimal test design procedure depends on a sufficiently large calibrated item bank. A study by Wainer, Kaplan, & Lewis (1992) indicated that the value of an objective function for a linear testlet constructed from a large item bank compares favorably to that of an adaptive test. Although the above described relation suggests that a comparison among different test types with one of the criteria in $\Phi(\cdot)$ seems possible, it should be emphasized that such a comparison is not straightforward. The function $\Phi(\cdot)$ is generally model-dependent and will have different ranges for different designs. This means that the criterion function values based on different numbers of θ parameters are not generally comparable.

References

Adema, J. J. (1990). *Models and algorithms for the construction of achievement tests.* Doctoral dissertation, University of Twente, Enschede.

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

Baker, F. B., Cohen, A. S., & Barmish, B. R. (1988). Item characteristics of tests constructed by linear programming. *Applied Psychological Measurement, 12,* 189–199.

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). 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. 200–208). New York: Springer-Verlag.

Berger, M. P. F. (1992b). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika, 57,* 521–538.

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics, 19,* 43–56.

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.

Berger, M. P. F., & Veerkamp, W. J. J. (in press). A review of selection methods for optimal test design. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 3). Norwood NJ: Ablex.

Boekkooi-Timminga, E. (1989). *Models for computerized test construction.* Doctoral dissertation, University of Twente.

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

Ford, I., Kitsos, C. P., & Titterington, D. M. (1989). Recent advances in nonlinear experimental design. *Technometrics, 31,* 49–60.

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.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993) Influence of item parameter errors in test development. *Journal of Educational Measurement, 30,* 143–156.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31,* 3–31.

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

Silvey, S. D. (1980). *Optimal design.* London: Chapman & Hall.

Theunissen, T. J. J. M. (1986). Binary programming and test design. *Psychometrika, 50,* 411–420.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103–135). Hillsdale NJ: Erlbaum.

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

Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 53,* 237–247.

Veerkamp, W. J. J., & Berger, M. P. F. (1993). *Some new item selection criteria for adaptive testing* (Research Report). Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement, 29,* 243–251.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1–14.

Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 24–35). Washington DC: United States Civil Service Commission.

Weiss, D. J. (1978). *Proceedings of the 1977 Comput-*

erized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Wu, C. F. J. (1985). Asymptotic inference from sequential design in nonlinear situation. *Biometrika, 72,* 533–558.

Wu, C. F. J., & 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.

Acknowledgments

The author expresses his appreciation to Wim J. Van der Linden for suggesting improvements in the text of an earlier version and to the reviewers for their comments.

Author's Address

Send requests for reprints or further information to Martijn P. F. Berger, Educational Measurement and Data Analysis, Department of Education, University of Twente, P.O. Box 217, AE Enschede, The Netherlands. E-mail: berger@edte.utwente.nl.