1

Optimal Design in Educational Testing

Steven Buyske

Rutgers University, Department of Statistics, 110 Frelinghuysen Rd, Piscataway, NJ 08854-8019, USA

1.1 Introduction

Formal job testing of individuals goes back more than 3000 years, while formal written tests in education go back some 500 years. Although the earliest paper on optimal design in statistics appeared at about the same time as multiple choice tests appeared, at the beginning of the twentieth century, optimal design theory was first applied to issues arising in standardized testing 40 years ago.

Van der Linden and Hambleton (1997b) suggest thinking of a test as a collection of small experiments (that is, the questions, or items) for which the observations are the test-taker’s responses. These observations allow one to infer a measurement of the test-taker’s proficiency in the subject of the test. As with most experimental settings, the application of optimal design principles can offer great gains in efficiency, most obviously in shorter tests. Since the cost of producing items can easily exceed US$100 per item, more efficient testing can lead to substantial savings.

The theory underlying most of modern testing is known as item response theory (IRT). In contrast to traditional test theory, IRT considers individual test items, rather than the entire test, to be the fundamental unit. It assumes the existence of an unobserved, or latent, underlying trait for both the proficiency of the test-taker and
for the difficulty of the individual item. The difference between the two, as well as other characteristics of the item, determine the probability that the test-taker will answer the item correctly.

### 1.1.1 Paper-and-pencil or computerized adaptive testing

Traditionally, standardized educational testing has been conducted in large-scale paper-and-pencil administrations of fixed-form tests. For example, in the United States some 3 million students take the SAT I and II tests on seven separate dates annually. These administrations feature a large number of students taking a small number of distinct, essentially equivalent, test forms. After the administration, both test-taker and item parameters are estimated simultaneously.

Although fixed-form tests can also be administered by computer, in recent years the leading alternative to paper-and-pencil testing has been computerized adaptive testing (CAT). In a CAT administration, a test-taker works at a computer. Because each item can be scored as quickly as the answer is recorded, the computer can adaptively select items to suit the examinee. The idea is that by avoiding items that are too hard or too easy for the examinee, a high-quality estimate of the examinee’s proficiency can be made using as few as half as many items than in a fixed-form test. CAT administrations can be ongoing. In the United States some 350,000 students take the Graduate Record Examination over more than 200 possible test days annually. Because of the need for on-line proficiency estimation, the item parameters are estimated as part of earlier administrations, known as item calibration, and so CAT is heavily dependent on efficient prior estimation of item parameters. Such testing is not limited to an educational setting; the US military and companies such as Oracle and Microsoft use CAT. Wainer (2000) gives a complete introduction to the subject, while Sands et al. (1997) and Parshall et al. (2002) give details on the implementation of computer-based testing.

### 1.1.2 Dichotomous response

The simplest IRT models apply when the answer is dichotomous: either right or wrong. By far the most common model for this situation are the 1-, 2- and 3-parameter logistic models (1-PL, 2-PL and 3-PL). The number of parameters refers to the parameters needed to describe each item. In the 3-PL model, the probability that a test-taker with proficiency \( \theta \) correctly answers an item with parameters \((a, b, c)\) is

\[
P(\theta \mid a, b, c) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}},
\]

(1.1)

where \( a \in (0, \infty) \), \( b \in (-\infty, \infty) \), and \( c \in [0, 1] \). Typical ranges in practice might be \( a \in [0.3, 3] \), \( b \in [-3, 3] \) and \( c \in [0, 0.5] \). The \( c \) parameter is often known as the
The 2-PL model is simply the 3-PL model with the pseudo-guessing parameter \( c \) fixed at zero. For the 1-PL model, \( c \) is fixed at zero and \( a \) is also fixed. The well-known Rasch model is equivalent to the 1-PL model. In all three cases, a scaling factor of \( D = 1.7 \) is often included in front of the \( a \) term to approximate the normal ogive. A brief, historically based introduction to dichotomous IRT models can be found in Thissen and Orlando (2001). Lord (1980) and Hambleton and Swaminathan (1985) have book-length treatments.
1.1.3 Polytomous response

There has been substantial recent interest in modelling all of the individual possible responses to an item, rather than just whether the response is correct or not (see, for example, van der Linden and Hambleton, 1997a; Thissen et al., 2001). There are a variety of models for polytomous response. We consider the nominal response model, first put forward by Bock (1972), as it includes several of the others as special cases (see Thissen and Steinberg, 1986). In the statistics literature it is also sometimes known as the multinomial logistic model. The nominal response model can be thought of as a combination of 2-PL models, one for each possible response to an item. If there are \( m + 1 \) possible responses, then the probability of a test-taker with proficiency \( \theta \) choosing response \( u \) is

\[
P_u(\theta \mid a, b) = \frac{\exp(a_u \theta + b_u)}{\sum_{v=0}^{m} \exp(a_v \theta + b_v)}.
\]

Because this model is underidentified, a constraint of either \( \sum_v a_v = \sum_v b_v = 0 \) or \( a_0 = b_0 = 0 \) is usually added. The vector pair \( (a, b) = (\{a_1, \ldots, a_m\}', \{b_1, \ldots, b_m\}') \) thus characterizes the psychometric properties of the item. Figure 1.2 shows graphs of \( P_u \) for a single item with three possible responses.

![Figure 1.2](image)
1.1.4 Information functions

For future reference it will be useful to write the Fisher information functions for proficiency and for the item parameters. The information for $\theta$ from a dichotomous item with parameters $(a, b, c)$ is

$$I_{a,b,c}(\theta) = \frac{P'(\theta)^2}{P(\theta)Q(\theta)} = \frac{a^2(P(\theta) - c)^2Q(\theta)}{(1 - c)^2P(\theta)},$$ (1.3)

where $P(\theta)$ is the probability of a correct response, as in Equation (1.1), and $Q(\theta) = 1 - P(\theta)$ is the probability of an incorrect response. Note that when $c = 0$, $I_{a,b}(\theta) = a^2P(\theta)Q(\theta)$, which is maximized when $P(\theta) = Q(\theta) = 0.5$, or when $b = \theta$. Figure 1.3 shows information functions for the three items displayed in Figure 1.1.

![Figure 1.3](image)

**Figure 1.3** Information functions for $\theta$ for three dichotomous items. When $c \neq 0$ the function is asymmetric

The information matrix for an item with parameters $(a, b, c)$ from a test-taker with proficiency $\theta$ is, writing $\bar{P}$ for $P - c$,

$$I_\theta(a, b, c) = \frac{Q(\theta)}{(1 - c)^2P(\theta)} \begin{pmatrix} (\theta - b)^2\bar{P}^2(\theta) & -a(\theta - b)\bar{P}^2(\theta) & (\theta - b)\bar{P}(\theta) \\ a^2\bar{P}^2(\theta) & -a\bar{P}(\theta) & 1 \end{pmatrix}.$$ (1.4)
For the nominal response model, the information for $\theta$ from an $(m + 1)$-choice item with parameters $(a, b)$ is

$$I_{a,b}(\theta) = \sum_{i,j = 0}^{m} a_i a_j (\delta_{ij} - P_i) P_j,$$

where $P_i$ is the probability of responding with choice $i$, and $\delta_{ij} = 1$ when $i = j$ and is zero otherwise. Notice that in the case of a dichotomous $(m = 1)$ item, this function reduces to the 2-PL information function for $\theta$ above. The information matrix for the item from a test-taker with proficiency $\theta$ has the form

$$\begin{pmatrix} A & C \\ C' & B \end{pmatrix}$$

where $A, B$ and $C$ are $m \times m$ matrix with entries of the form

$$A_{ij} = (\theta - b_i)(\theta - b_j)(\delta_{ij} - P_i) P_j,$$

$$B_{ij} = a_i a_j (\delta_{ij} - P_i) P_j,$$

$$C_{ij} = -(\theta - b_i) a_j (\delta_{ij} - P_i) P_j.$$

Notice again that in the case of a dichotomous item, this function reduces to the 2-PL information function for $(a, b)$ above. The information function for $\theta$ for the polytomous item in Figure 1.2 is shown in Figure 1.4.

![Figure 1.4](attachment:figure14.png)  
**Figure 1.4** Item information for $\theta$ for a polytomous item
1.1.5 Design problems

Design problems in educational testing are of two types. The first type, known as test design, has to do with the optimal selection of items for proficiency estimation. The second type, sometimes called the sampling design problem, has to do with the sampling of test-takers for optimal estimation of the item parameters.

There are two important aspects to optimal design in educational testing. The first is that the model is nonlinear. In a linear optimal design problem, the question of how to sample the levels of the independent variables to estimate the model parameters efficiently depends on a function of Fisher’s information matrix for the model parameters. In the nonlinear case the Fisher information matrix will also depend in an essential way on the model parameters. This dependence has two consequences. First, the design will be optimal only locally, in the sense of only for the specific model parameter values. Other values will lead to different designs. Second, as the model parameters are themselves unknown, it is not a simple matter to find even a locally optimal design. The most common approach is to work sequentially. As the parameter estimates change, the design changes with them. An alternative is to use Bayesian methods. A reasonable prior for the parameters can be integrated away, eliminating the design dependence on the specific values of the parameters.

The second important aspect is that the underlying model for IRT is a latent variable model. Whether one is creating a design for estimating proficiencies or item parameters, one cannot exactly specify the design levels. They are latent variables, and cannot be directly observed. They can, however, be estimated, and one can work with those estimates. Even with the estimates, however, one cannot specify an observation at a specific value. In particular, a design may call for an item to be tested on an examinee with a particular value of $\theta$, but one should not expect that such an examinee will be available. Instead, one must work with random samples.

1.2 Test Design

Birnbaum (1968) first proposed the modern conceptual procedure for test construction. First, one determines the purpose of the test. Second, the purpose of the test determines the target for the test information function (TIF). The function TIF($\theta$) is the expected Fisher information from the test for an examinee with proficiency $\theta$. For an aptitude test the ideal TIF might be uniform across a wide range of proficiencies, while for a certification test the ideal TIF might be high near the certification boundary and lower away from the boundary. The third step in test assembly is the selection of items so that the actual TIF, the sum of the item information functions, comes close to the target, subject to a variety of constraints discussed below. A thorough review of test design is given in van der Linden (1998), which introduces a special issue of *Applied Psychological Measurement* devoted to test construction issues.
1.2.1 Fixed-form test design

Fixed-form test design problems can be classified as unconstrained or constrained. Unconstrained problems consider optimal designs in the purely psychometric IRT setting, where there are no external constraints as to the items selected. Constrained problems incorporate these constraints, which may force a balance among items of a certain type, limit the total word length of items, or restrict the use of too similar items. Solutions to constrained problems have cast the problem as a 0–1 linear programming problem, as a problem in network flow programming, or used less formal heuristic approaches. Readers not familiar with optimization techniques may wish to consult texts such as Nocedal and Wright (1999) and Ahuja et al. (1993).

Berger (1998) reviews the use of optimal design theory to design tests. Suppose that instead of starting with an ideal test information function one simply wanted to efficiently estimate the \( \theta \)'s for a population of test-takers. Begin with a vector \( \theta = (\theta_1, \ldots, \theta_k)^t \) of target proficiency values and their weights \( W \) corresponding to a discrete proficiency distribution. The set of possible item vector pairs \( (a, b) \) or triplets \( (a, b, c) \) forms the design space. The optimal test design problem can then be stated as a problem of maximizing a function of the Fisher information for the vector \( \theta \) with weights \( W \) over the design space \( \{(a, b)\} \) or \( \{(a, b, c)\} \). This design space may be constrained by practical issues. \( D \)-optimality seems a natural design criterion, or one could use the maximin criterion, which maximizes the minimum information over \( \theta_i \) (van der Linden and Boekkooi-Timminga, 1989). Berger (1998) reports that for very high discrimination items, the optimal designs feature difficulties roughly following the distribution of the test-takers’ \( \theta \)'s. For more moderate discrimination items, and a normal distribution of \( \theta \), both the \( D \)-optimal and maximin-optimal designs place the item difficulties all close to the centre of the distribution of \( \theta \). For a uniform distribution of \( \theta \), \( D \)-optimality leads to a bimodal distribution and maximin leads to a trimodal distribution. Berger suggests this feature arises because the uniform distribution is trimmed and relatively high information is thus needed near the boundaries. Low discrimination items have fairly flat information, so that for a normal distribution of \( \theta \) setting the difficulties close to the centre gives the most information.

Berger (1994b) showed how test design for fixed-form tests, adaptive tests and testlets can all be placed into the same optimal design framework. Testlets are collections of items developed as a small test with the idea of simplifying some aspects of adaptive testing. Berger showed that optimal design could be used to select the items for the testlet, called within-testlet optimality, and to select testlets to construct a test, called between-testlet optimality. In all cases, a sequential block (with blocks possibly of size one) design is used. In an adaptive test, the \( \theta \) estimate changes with each step; in a fixed-form test, the \( \theta \) vector and weights remain the same throughout.

Berger and Mathijsse (1997) have examined optimal test design for the nominal response IRT model. A new difficulty that arises in this context is that every
category in an item has its own parameters that determine its relative position, within the item, on the difficulty scale. A test design will thus depend on the choice of the number of categories per item and their parameters.

An interesting result is that the maximum information for a given \( \theta \) is obtained by an item for which the responses with the lowest and highest discrimination parameter \( a_i \) have a probability of selection of 0.5; that is, the item essentially reduces to an optimal dichotomous 2-PL item. In applying the \( D \)-optimality criterion, Berger and Mathijssen looked at item location vectors \( b \) for fixed discrimination vectors \( a \) for three and for four category items. As with the dichotomous model case, they found that for a normally distributed sample of examinees, the \( D \)-optimal test design required a single location vector, while for a uniform distribution of examinees a wider spread of location vectors was called for. As the category discrimination parameters increased, so did the location parameters.

Mathematical programming approaches to constrained test design appear to have been first put forth by Votaw (1952) and Feuerman and Weiss (1973). Although the textbook description of test construction as matching the test information function to the target sounds straightforward, in reality test construction includes a thicket of constraints. For example, there may be constraints on the number of items in the test, minimum and maximum number of items for specific subject areas or item types, and dependency constraints that if certain items are included others cannot be (because of overlap, or because one item gives clues to others) or must be (several items about the same reading passage, for example). A comprehensive list of constraints is given in Linden and Boekkooi-Timminga (1989).

To cast test construction as a 0–1 linear programming problem, let \( x_i, i = 1 \ldots, M \) be a 0–1 variable indicating whether item \( i \) is included in the test out of the \( M \) items in the test bank. If the test construction objective is to maximize information at \( \theta_0 \), then the function to be maximized is \( \sum_{i=1}^{M} x_i I_i(\theta_0) \), subject to constraints on the \( x_i \). For example, the requirement of a test of \( k \) items is simply \( \sum_i x_i = k \), while a requirement of a minimum of \( j \) items of a certain type is \( \sum_{i \in T} x_i \geq j \), where \( T \) is the set of items of that type. If the \( x_i \) variables were not constrained to be 0–1 variables, the resulting constrained optimization problem would be easily solved using linear programming. With the 0–1 constraint, the problem is known to be NP-hard, meaning the problem lies in the class of problems whose solution time is believed not to be bounded by a polynomial in the size of the problem. However, optimal solutions are not really required, just ‘very good’ ones. Approximate solutions can be found by limiting the search time of the branch-and-bound search or by using optimal rounding, and can be reasonably fast. The 0–1 programming approach was formulated by Theunissen (1985, 1986). Later work developing the concept can be found in Boekkooi-Timminga (1987, 1990), van der Linden and Boekkooi-Timminga (1989), Adema \textit{et al.} (1991), Adema (1992, 1998), van der Linden and Adema (1998) and van der Linden (2000).

The second mathematical programming approach to test construction uses network-flow programming (Armstrong \textit{et al.}, 1992). Here one conceptualizes
supply nodes, representing items; demand nodes, representing constraints such as a minimum and maximum number of items of a certain type; directed arcs, representing the flow of items from demand nodes to supply nodes; and the cost of flow, which can be used to represent the item information. Network flow problems can be efficiently solved using the simplex algorithm. The principal advantages of a network-flow formulation of test construction are that optimal solutions can be found extremely quickly, and these solutions naturally involve only 0–1 variables. The principal disadvantage is that certain constraints, such as quantitative constraints (total word count, for example) and dependency constraints do not easily fit into the network formulation. Lagrangian relaxation is a technique to bring such constraints into the objective function by multiplying the constraint by a Lagrange multiplier and adding it to the objective function. By systematically varying the multiplier and solving the new network flow problem, near-optimal results can be very quickly found. Armstrong et al. (1995, 1996, 1998a, b) show how to use Lagrangian relaxation and heuristic search along with network flow to solve test construction problems with these non-network constraints.

In both mathematical programming approaches other objective functions may be more desirable than the information at a fixed value $\theta_0$, and can be incorporated into the 0–1 linear programming model. For example, one can minimize the discrepancy between the test information function and the target over a discrete set of $\theta$ values. Van der Linden and Boekkooi-Timminga (1989) minimized the discrepancy from the shape of the target test information function, arguing that the actual values of the test information carry little meaning for test designers. Armstrong et al. (1998b) used classical reliability as an objective function.

In the mathematical programming paradigm it is important that a feasible solution exists; that is, that a test does exist satisfying all of the constraints. Timminga (1998) discusses diagnosing and repairing an infeasible model. It is also possible to treat the constraints as desirable conditions that can be violated if necessary. The heuristic method of test assembly now known as the weighted deviations model (Luecht and Hirsch, 1992; Swanson and Stocking, 1993; Stocking and Swanson, 1993; Luecht, 1998) incorporates constraints into the objective function by penalizing deviations from the constraints. The test is constructed using the greedy algorithm: the best item is selected, the objective function is regenerated to incorporate the prior decision, a new best item is added to the test, and the process repeats. Because constraints are incorporated into the single objective function, weights can be used to emphasize certain constraints over others. The greedy algorithm assures rapid test assembly but because the test is assembled sequentially there is no guarantee of overall optimality. Optimal design criteria could be brought into the weighted deviation model by using an optimal design criterion as the basic objective function before incorporating the constraints. Although algorithmically less sophisticated than other constrained optimization approaches, the weighted deviations model is widely used because of its speed, flexibility, ease of interpretation, and its guarantee of always finding a solution.
1.2.2 Test design for CAT

In CAT, the testing algorithm has the advantage of having an estimate of the test-taker’s proficiency at each step, which leads naturally to a (constrained) sequential optimization problem. The original approach to item selection in a CAT was, at step $k$, to treat the current estimate $\hat{\theta}_k$ as the true value of $\theta$ and to select next the item with the highest Fisher information for $\hat{\theta}_k$. Chang and Ying (2004) showed that under the ideal conditions of an infinite item pool this design is asymptotically optimal for the 1-PL model, and with certain restrictions, principally bounds on $a$ and $c$, is also asymptotically optimal for the 2-PL and 3-PL models.

Near the start of the exam, however, the estimate of $\theta$ may be quite poor. Selecting items by maximum information tends to use high discrimination items early, although these items may contain little information if $b - \theta$ is actually large (see Figure 1.2). Veerkamp and Berger (1997) thus looked at an item selection rule based on the Fisher information integrated over an interval of $\theta$-values, using various weights. With a delta function at $\hat{\theta}_k$, one recovers the traditional maximum Fisher information rule. They also considered uniform weights over a confidence interval for $\theta$ and weights proportional to the likelihood function for $\theta$. They found that weighting by the likelihood gave slightly superior results to either maximum information or the information integrated over a confidence interval. Chang and Ying (1996) proposed using an integrated Kullback–Leibler information criterion for item selection, again as protection against poorly estimated $\theta$ near the beginning of the test. Chen et al. (2000) compared all these item selection rules with simulations and, although they found differences in short tests of less than 10 items, found no substantive differences in longer tests.

Stocking and Swanson (1993) adapted the weighted deviations method to handle item constraints in CAT. Their method is based on selecting the maximum information item as modified by constraints, but any of the selection rules in the preceding paragraph could be used prior to the constraint modification.

A practical constraint in CAT goes by the name ‘exposure control’. Because a test in the CAT format is generally offered frequently, there is a distinct possibility that certain items can become known in advance to certain test-takers. Unexpectedly high exposure of certain items lead to considerable embarrassment in the early administration of the CAT version of the GRE exam (Wainer and Einor, 2000). Less dramatic is the possibility of certain items being underutilized, leading to inefficiency and increased cost. The most popular approaches are based on work by Sympson and Hetter (1985) and assign selection probabilities to each item. While widespread, such methods have the disadvantage of requiring extensive simulation to assign the probabilities. A review of probabilistic methods of controlling item exposure is given in Stocking and Lewis (2000). These methods can be incorporated into the weighted deviations method and other methods of test construction.

An alternative to probabilistic methods was suggested in Chang and Ying (1999) and Chang et al. (2001). They propose stratifying the item pool by the $a$-level. Low
discrimination items, with their flatter information curves, would be used early, when the proficiency is poorly estimated, and higher discrimination items, with more concentrated information, would be saved for the later stages of the test when the proficiency is better estimated. Item exposure is controlled because low discrimination items are forced into use, thus reducing the skewness of the item exposure distribution.

The speed of modern computing has enabled a promising approach to on-the-fly construction of adaptive tests satisfying the various test constraints. Van der Linden and Reese (1998) show that at each stage of an adaptive test an algorithm can select the next item by assembling a full test, fixing the items already administered. From this full test, a single next item is selected and the unselected items are returned to the pool. Following the test-taker’s response, the process is repeated. Since each full test satisfies the constraints, the actual adaptive test automatically satisfies the constraints as well. Although Van der Linden and Reese used 0–1 programming in their paper, presumably the technique would apply to any method of test construction. Although computationally intensive and algorithmically sophisticated, the flexibility and potential generality of this technique make it likely to see wide use in the future.

Because item selection uses the estimate \((\hat{a}, \hat{b}, \hat{c})\) as the true parameter value, there is a potential for serious errors in CAT due to capitalization on chance. Hambleton et al. (1993) and van der Linden and Glas (2000a) discuss the phenomenon and outline approaches to reduce its impact.

1.3 Sampling Design

We now reverse perspective and look at designs to estimate item parameters efficiently. The problem is in some ways simpler than test design, because there are no test design constraints, yet it does involve more parameters to be estimated. The problem differs depending on whether the item calibration is to take place in a paper-and-pencil or a CAT format. In the former, at most one can specify the rough distribution of the proficiencies of the test-takers. In a CAT format, a design algorithm can take advantage of the current estimate of the test-taker’s proficiency to obtain a more efficient design.

1.3.1 Paper-and-pencil calibration

Originally, items were calibrated using test-taker samples similar to the test-taker population. Lord (1962), Pandey and Carlson (1976) and Lord and Wingersky (1985) were able to show that greater precision in item parameter estimates could be obtained if the calibration sample were explicitly chosen for that purpose. In particular, they showed that giving each item to an appropriately matched subsample of test-takers was more efficient that using the entire sample for the
entire set of items. Stocking (1990) considered the effect of different sampling schemes on the information for each item parameter separately in 1-PL, 2-PL and 3-PL models, showing, parameter by parameter, how to interpret the results of Wingersky and Lord (1984) that a rectangular sample of proficiencies worked well in the 3-PL model for simultaneous estimation of proficiency and item parameters.

Berger (1991) brought formal optimal design theory to the question by investigating sampling designs using the $D$-optimality criterion for 1-, 2-, and 3-PL models. First considering designs with only a single sample, he showed that for the 1-PL model the design efficiency drops as the standard deviation increases, because fewer examinees will have $\theta$’s near the maximally informative value of $\theta = b$. For the 2-PL model, the efficiency first increases with the standard deviation, because $a$ can be more accurately estimated with a spread sample, and then decreases. The value of the standard deviation giving maximum efficiency is a decreasing function of $a$, since for high values of $a$ even $\theta$’s moderately near to $b$ will give values of $P$ near 0 or 1, and so have little information for either $a$ or $b$. Because of the need to estimate $c$, for 3-PL models the efficiency increases with the standard deviation of the sample.

Berger also explored alternative sampling designs to roughly match $\theta$’s and $b$’s. For the 1- and 2-PL models, using a moderate number of partitions of the test-taker sample and partial test subsets roughly matching the proficiencies of the subsamples gave much higher efficiencies than giving the entire test to the full sample. The efficiency improvement depends on the discrimination: higher discrimination items show greater increases in efficiency. Except for the easiest items, in a 3-PL model the greatest efficiency comes from using the full sample.

In the 1-PL case, it is easy to show that the information for $b$ is maximal if all test-takers are selected so that $\theta = b$. The optimal sampling of test-takers was handled by Berger (1992) for $D$-optimality in the 2-PL case. The optimal design for a given item has equal weights at two design points, namely $\theta_1$ and $\theta_2$, such that $P(\theta_1) = 0.176$ and $P(\theta_2) = 0.824$ (White, 1975; Ford, 1976). Explicitly, $\theta_i = b \pm 1.5434/a$. Again we see that as $a$ increases the points should be closer to $b$. (For $A$-optimality, the results are not so simple, as the best design depends heavily on $a$ – the optimal distance between the points decreases with $a$.) These results imply that bimodal calibration samples will have higher efficiency than a single normal sample. Samples with a standard deviation between 1 and 2, with a gap between the modes and $b$ of about $1.5/a$, can be expected to be 60–80% efficient compared to an ideal sample. This means, for example, that an exam for 5th graders would be more efficiently calibrated (on the order of 50% more efficient) by a sample consisting of half 4th graders and half 6th graders than by a sample consisting fully of 5th graders. Berger also showed how the method of Wynn (1970) can be adapted to determine locally $D$-optimal sampling designs for sets of test items. The greatest gains in efficiencies, as compared to a $N(0,1)$ sample, were seen in tests with low discriminating items and a wide range of difficulties. (This sort of test is perhaps the easiest to construct; creating items with
high values of $a$ is a surprisingly difficult task.) The typical sample design is essentially trimodal, with modes at the extremes and the centre of the range for $\theta$. A review of optimal sampling designs for paper-and-pencil calibration can be found in Berger and van der Linden (1992).

$D$-optimal designs for a range of polytomous models were considered by Holman and Berger (2001), who found that $D$-optimal designs can reduce the calibration sample size by up to 30%.

### 1.3.2 CAT calibration

When an item is to be calibrated in a CAT administration, there is the possibility of a substantial efficiency gain by administering an item only to those test-takers who will add the most information to the item parameter estimates. On-line calibration easily lends itself to the sequential procedure of adapting the calibration sample as the item estimates evolve. Berger (1994a) and Jones and Jin (1994) first demonstrated this sequential $D$-optimal design procedure for the 2-PL model. Berger showed that there is very little loss in efficiency in making the procedure block-sequential rather than sequential. Jones and Jin demonstrated in simulations that the sequential approach yielded designs with about 70–80% of the efficiency of optimal design for known item parameters. Simple normal sampling gave efficiencies ranging from 0.75, for a proficiency equal to the sample mean, down to 0.01 for high discriminating items with extreme difficulties. As CAT administrations require a steady stream of newly calibrated items while reducing the number of items any individual test-taker will take part in calibrating, high efficiency in item calibration is quite desirable.

Buyske (1998a, b) took a different approach to a criterion function. His criterion is based on minimizing the information loss for $\theta$ that will occur due to estimation errors in the item parameters. For the 2-PL model the criterion takes the form

$$
\frac{1}{a^2} \text{Var}(\hat{a}) + \text{const} \times a^2 \text{Var}(\hat{b})
$$

where the constant depends on the expected distribution of proficiencies when the item is used in production. In simulations these designs were 50% more efficient (in the sense of the calibration sample size needed for equally good proficiency estimation) than a standard normal distribution and 10% more efficient than a $D$-optimal design. In the 2-PL case the optimal design points are moved in slightly from $D$-optimal case, from $\theta$'s giving probabilities $P$ near 0.18 and 0.82 for $D$-optimality to 0.25 and 0.75 for this revised criterion. A similar contraction of design points is found for the 3-PL model.

In practice, one has more than one item to calibrate at a time, and has a stream of test-takers who must be given calibration items. Thus we again have a constrained nonlinear optimization problem. This situation was first investigated by Jones et al.
(1997) for the 2-PL model. The problem is: given $m$ items to be calibrated, and $n$ test-takers, match test-takers with items to maximize information. An additional constraint is that every test-taker should receive the same number of items to calibrate and each item is calibrated by the same number of test-takers. By casting the problem as a network-flow problem and using a gradient-search algorithm, the authors were able to find exact designs quickly and with efficiency about 50% greater than random allocation. As in the two papers by Berger (1992, 1994a), the designs tend to have proficiencies cluster around the difficulty level of the items, with tighter concentrations for higher discriminating items. The same problem was investigated for the 3-PL model by Jones et al. (1999).

Testlets pose a slightly different calibration problem. Jones and Nediak (2004) report on $D$-optimality-based sequential calibration of testlets. They examined both constrained and unconstrained methods. They found that about 30% more items calibrated by optimal design than by a random sample were acceptably close to the true item characteristics, as measured by the weighted Hellinger deviance, a measure of the distance between the item response functions of the estimated and the true items.

$D$-optimal solutions to item calibration designs are in fact locally optimal designs, because of the dependence of the information matrix on the item parameters. Berger et al. (2000), working with the nominal response model, also addressed the issue of designs for calibrating items using a minimax $D$-optimality criterion. The minimax approach avoids the difficulty of local optimality by minimizing the worst performance of the item calibration, as measured by the $D$-optimal criterion, over the prespecified region of item parameters. The design points determined by this criterion are more spread out than points determined by the $D$-optimality criterion, yet have high efficiency relative to the local $D$-optimal design. Because the design does not require estimation of the item parameters, by its very nature it is more robust against misspecification of those values.

1.4 Future Directions

At this time the development of optimal test and sampling design theory for purely psychometric IRT seems mature for both dichotomous and many polytomous models. Much of the recent interest has been in the practical issues that arise in applied testing. The trend in increasing computation speed and algorithmic improvements means that constrained nonlinear optimization techniques such as 0–1 programming or network-flow algorithms will become practical competitors to the weighted deviations algorithm in test design. The on-the-fly test construction method for CAT developed in van der Linden and Reese (1998) also seems likely to see future development and wide use.

Exposure control remains an important issue in CAT, both because of the risk of cheating and the cost of developing underutilized items. A conceptual breakthrough is still needed here. Perhaps the way lies in thinking of CAT test construction not as
a constrained sequential optimization problem for each test-taker, but in thinking of it as a sequential optimization problem for the overall pool of test-takers, with both the usual constraints on each test as well as constraints on the item pool limiting each item’s exposure.

The very efficiency of optimal designs for online item calibration has led to some difficulties. Model-fit diagnostics for items such as multidimensionality tests and differential item functioning (different item characteristics in different groups) rely on large samples for their power. Calibration designs to date have not incorporated the need for efficient item diagnostic testing.

Many of the future challenges in optimal design theory in educational testing will follow future trends in testing itself. There is almost no design work in alternative IRT such as multidimensional traits or non-parametric models. The use of computer-generated items (item cloning) as well as innovative items, such as items requiring constructed responses with automated scoring, is gathering increased interest. As these concepts move from journal articles to actual tests there will be an increased and pressing need for effective designs incorporating them.

Acknowledgements

The author would like to thank the editors and two anonymous referees for their valuable comments.

References


REFERENCES


