Graduate Management Admission Council®

Up-and-Down Procedures for Approximating Optimal Designs Using Person-Response Functions

Yanyan Sheng

Southern Illinois University

Nancy Flournoy and Steven J. Osterlind

University of Missouri, Columbia

Presented at the IRT Estimation and Polytomous CAT Paper Session, June 7, 2007

Abstract

Algorithms developed for item selection in computerized adaptive testing are exclusively from the item response theory perspective, where responses are examined at the item level. They are limited when all item characteristics are not readily known. This paper illustrates an item selection algorithm using person response functions. When targeting a single percentile, median difficulty level can be an estimate of the person's ability level. Often however, estimating more than one percentile or the entire person response function is desirable. This study considered optimal designs that either minimize the variance of one percentile estimate or a function of the variance-covariance ellipse jointly for all parameter estimates. Due to the problem that an optimal design for nonlinear models depends on unknown parameters, sequential up-and-down designs for approximating optimal designs are considered. Examples are provided to illustrate the efficiency of the procedures.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Sheng, Y., Flournoy, N., & Osterlind, S. J. (2007). Up-and-down procedures for approximating optimal designs using person-response functions. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Yanyan Sheng, Department of Educational Psychology & Special Education, Southern Illinois University, Carbondale IL U.S.A. ysheng@siu.edu

Up-and-Down Procedures for Approximating Optimal Test Designs Using Person Response Functions

In recent decades, the popularity of computerized adaptive testing (CAT) in educational and psychological test administrations has increased dramatically. Although CAT can be contrasted with test administration in traditional modes (e.g., paper and pencil), it is actually much more than mere presentation of assessment stimuli by computer. Traditional assessment presentations are typically "flat," where all examinees respond to all items and scores are calculated using a simple, summing model, although many types of scores can be derived. In contrast, CAT typically uses a maximum likelihood procedure to sequentially approximate an examinee's ability or trait level during the assessment, based solely upon the number of items presented and the examinee's individual pattern of correct endorsements (Wainer et al., 2000). After some initial "anchoring" to get the process started, each newly-presented item provides more data for another sequential, and more accurate, estimation. Thus, in CAT the "test" is not a single set of items that all examinees respond to but, rather, is composed of items uniquely matched to an examinee's trait level, meaning that each individual responds to a different set of items (except examinees who have identical response patterns). Presumably, through successive approximations ever better trait estimates are made with each additional response. Eventually, the estimation asymptotes at some level; and, as next-presented test items yield only inconsequential improvements, the process stops. The final trait estimate is the examinee's reported score.

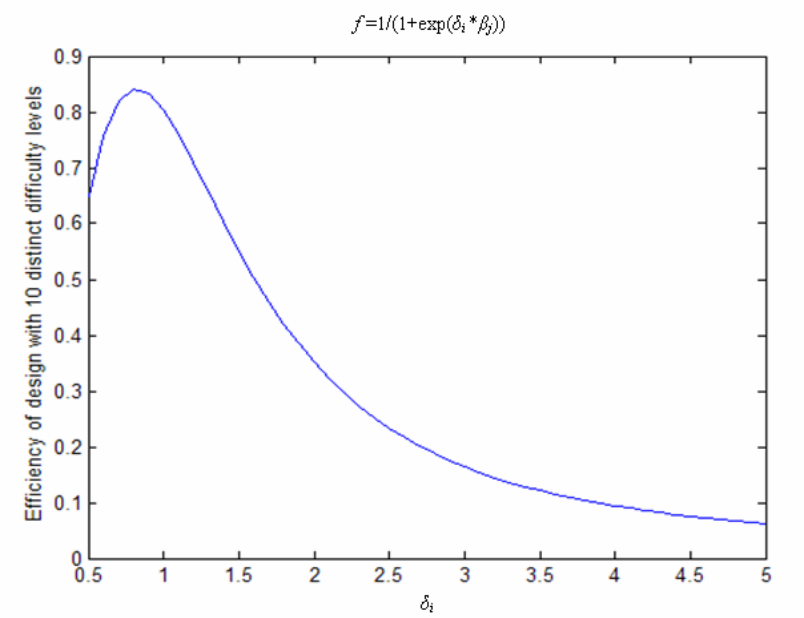
Algorithms developed for item selections in CAT such as the maximum information method (e.g., Thissen & Mislevy, 2000) are exclusively from the item response perspective, where responses are examined at the item level. They perform well only when the exact forms of item response functions (IRFs) are known. However, these algorithms are limited in situations where the item characteristics are not readily known. This paper introduces item selection from a person response perspective and focuses on using an up-and-down design to estimate an examinee's trait level. Up-and-down methods are mainly used by biologists for sensitivity testing. Lord (1970) first applied them in sequential designs for tailored testing and soon stopped using them, recommending other methods. Since then, up-and-down designs have not been of interest in the CAT literature because they are less efficient when the exact form of an item response function is known. They are, however, attractive for item selection from the person response perspective, especially when one knows only item difficulty levels. It can be shown that with this approach a trait (θ) estimate targets certain percentiles of interest in relation to the difficulty level of the overall test, such as the median difficulty level (parallel to ED50 in bioassay). When this is mapped as a function, it is termed the location of the person response function (PRF; Lumsden, 1978; Sijtsma & Meijer, 2001; Trabin & Weiss, 1983), or sometimes the threshold difficulty level. The adaptive design targeting only the median is usually taken to be the classical up-and-down design of Dixon & Mood (1948; von Békésy, 1947; cf. Lord, 1970; 1971).

Often, in educational measurement, estimating more than one percentile or the entire PRF is desired. To illustrate this, a test can easily be constructed with a set of arbitrarily chosen 10 (or some number) distinct difficulty levels. Nevertheless, such a design is found to be less efficient than optimal designs adopted in areas such as bioassay (e.g., see Figure 1, where the efficiency of the design is shown to be consistently smaller than that for the D-optimal design for persons with an average θ or with θ two standard deviations above the mean). This study considered optimal designs that either minimize the variance of one

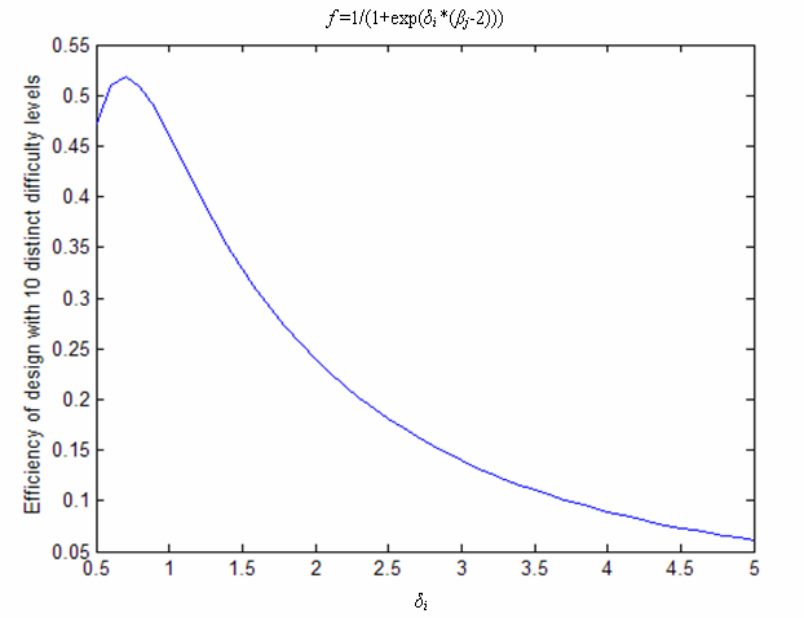
percentile estimate or a function of the variance-covariance ellipse jointly for all parameter estimates. The latter is more meaningful for measuring psychological variables.

Figure 1. The Efficiency [det(M^{-1})] of a Design With 10 Distinct Difficulty Levels Relative to the D-Optimal Design

a. $\theta_i = 0$



b. $\theta_i = 2$



A difficulty arises in formulating the PRFs because they cannot be directly known, as they are nonlinear and rely on unknown parameters. One way to overcome this problem is to use sequential designs to approximate optimal designs. This study focused on using

sequential up-and-down designs for the approximation of a range of response estimates. For estimating the median difficulty level (such as the ED50), the so called *c-optimal design* prescribes selecting items so as to minimize the variance of the estimator. In practice, this amounts to selecting all items at the median difficulty level. The c-optimal design can be approximated by the classical up-and-down test.

For estimating all the model parameters, the so called *D-optimal design* prescribes selecting items so as to minimize the determinant of the parameters' variance-covariance matrix. For PRFs, the D-optimal procedure is to select items at two percentiles not at the median difficulty level. The D-optimal procedure can be approximated by using two independent biased coin up-and-down procedures (Durham & Flournoy, 1994) that target the optimal percentiles.

It has to be noted that up-and-down designs do not rely on the form of the model and thus are nonparametric in nature. Regardless, to study the properties of the designs in appraising mental constructs, an underlying person response model was assumed in this study.

IRFs and PRFs

In modern measurement theory, IRFs are a class of probabilistic functions illustrating the interaction between examinees and test items. For the binary responses (0 and 1) of test item endorsement, denoted as Y, the parametric probability functions of correct responses are usually modeled by a logistic or by a normal ogive item response model. As an illustration, the two parameter logistic (2PL) IRF specifies that the probability of a correct answer to item j, denoted by $P_j(\theta)$ is

$$P_j(\theta) = P(Y_j = 1 | \theta) = \frac{1}{1 + \exp\{-\alpha_j(\theta - \beta_j)\}}, \quad (1)$$

Where α_j and β_j denote the discrimination and difficulty parameters for item j and θ denotes the continuous latent trait parameter scale, with θ_i being the trait level for person *i*. The IRF describes the probability of a correct response as a function of θ and fixed item parameters (e.g. α_j and β_j in the 2PL IRF). When the slope parameter α_j is assumed to be the same for all items, the logistic form of the model is known as the Rasch model

Analogous to the IRF, the PRF describes the probability of a correct response as a function of item difficulty and a fixed person parameter. Hence, what constitutes a person response must be defined appropriately for each person. Let Y_i be the random response variable for person *i* and let $Y_i = 1$ if the person has a correct response, and let $Y_i = 0$ otherwise. A continuous latent difficulty scale, denoted by β, is assumed, with β_j being the location of item *j* in the population of all the items. For a given person *i* with trait level θ_i, the PRF is defined by

$$P_i(\beta) = P(Y_i = 1 | \beta). \quad (2)$$

The probability function can be any non-increasing function of the difficulty scale parameter, β. In this study, the following logistic form was assumed:

$$P_i(\beta) = \frac{1}{1 + \exp\{-\delta_i(\theta_i - \beta)\}}, \quad (3)$$

where δ_i is the slope parameter for person *i*. δ_i can be interpreted as how different the person's probability is in correctly responding to easy and difficult items respectively and is assumed to vary from person to person.

Given that the *n*th item administered to person *i* has difficulty level $B(n) = \beta$, the probability of correct response is $P_i(\beta) = P[Y_i(n) = 1 | B(n) = \beta]$. The θ level that yields a target difficulty level Γ $(0 < \Gamma < 1)$ is defined as θ_Γ, so $\Gamma = P[Y_i(n) = 1 | B(n) = \theta_\Gamma]$. Under the logistic PRF, it can be shown that $ED100\Gamma = \theta_i + \gamma/(-\delta_i)$, where $\gamma = \log[\Gamma/(1-\Gamma)]$. Hence, the latent trait level, θ_i, is actually the median difficulty level (ED50) for person *i*.

The remainder of the paper is organized as follows. First, we describe optimal designs for educational and psychological measurement and discuss why they require successive approximations. Next, we review the classical up-and-down design and generalize it as needed to target arbitrary percentiles. Then, we show how the up-and-down design can be used to approximate the optimal designs. Finally, we give examples and evaluate their performance.

Optimal Designs

An optimal design is defined as a design that minimizes or maximizes some criterion function (Atkinson & Donev, 1996). Optimal designs emerged from the framework of design of experiments developed by Fisher (1935), with recent important developments in the literature, including Silvey (1980), Atkinson and Donev (1992), Pukelsheim (1993), and Fedorov and Hackl (1997). In the implementation of optimal designs in psychometric studies, all designs are considered approximate designs, which means that the designs are actual conditional probability measures. This suggests that a design consists of (1) a set of items with various difficulty levels, and (2) the assignment of a proportion of the fixed number of items to each difficulty level. Optimal (approximate) designs have been adopted to target the efficient estimation of item parameters (e.g., Berger ,1992; Berger, 1994; Buyske, 1998; Stocking, 1990, among others) as well as trait parameters (e.g., Berger & Mathijssen, 1997). Typically, these presentations focus on the 2PL IRF for parameter estimation. The focus here is on optimal trait estimation using PRFs.

For the logistic PRF specified in equation 3, in which items only differ in their difficulty levels, a design is denoted by ξ,

$$\xi = \begin{Bmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \\ w_1 & w_2 & \cdots & w_K \end{Bmatrix}. \qquad (4)$$

The vector $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_K]^T$ consists of *K* distinct difficulty levels and is referred to as a vector of optimal design points. The $\{ w_k \}$ are weights that indicate the proportion of times items of each difficulty level are presented to an examinee. It is assumed that $w_k \geq 0$ for *k* = 1, ..., K,

$$\sum_{k=1}^{K} w_k = 1$$

and $1 \leq K \leq N$, where *N* is the total number of items presented to the examinee. When *K* = *N* and $w_k = 1/N$ for all *K* levels, all items have different difficulties; on the other hand, if *k* = 1, then all *N* items have the same difficulty level. It is possible to view a design as a discrete probability measure with a probability distribution defined over the design space which, in

this case, is all items that are available in an item bank. In this study, an optimal item is an item with optimal values for the item difficulties.

If the proportion of correct responses in each of K classes or groups is given by $p = [p_1, p_2, \ldots, p_K,]T$, then even with dependent observations from up-and-down designs, the likelihood function of θ_i given $\boldsymbol{\beta}$ and $\mathbf{p}$ is

$$L = \prod_{k=1}^{K} F(\beta_j)^{w_j p_j} [1 - F(\beta_j)]^{w_j (1-p_j)} \tag{5}$$

(Rosenberger, Flournoy & Durham, 1997). The asymptotic efficiency of the maximum likelihood estimators of the person parameter θ is related to the Fisher information matrix $M(\theta_i \mid \boldsymbol{\beta}, \mathbf{W})$ defined below. The objective of using an optimal design for the estimation of θ parameters is to maximize information on the parameters. This can be done by selecting values for ξ that will maximize some function of the information contained by the information matrix.

White (1973) showed how the results of linear design theory can be adapted to nonlinear models, including binary regression models. The non-linearity of the person response relationship (as is shown in equations 3 and 4) implies that the measure of information is dependent on the unknown parameters so that one has to specify a best guess or approximate them through sequential procedures in order to construct an applicable design. In the literature, several local optimality criteria have been proposed that are defined on the information matrix. This study focused on the c-optimality and D-optimality criteria.

C-optimal Design

The c-optimal design minimizes the variance of the maximum likelihood estimate of the $100\,\Gamma$ percentile difficulty level (ED$100\,\Gamma$ in bioassay). The asymptotic variance of the median difficulty level (ED50), $\hat{\mu}_{.5}$ (which is equivalent to $\hat{\theta}_i$), can be written as:

$$\text{var}(\hat{\mu}_{.5}) = \mathbf{c}^T M^{-1}(\vartheta_i; \xi)\mathbf{c}, \tag{6}$$

where $\mathbf{c} = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$, $\vartheta_i = (\theta_i, \delta_i)^T$, and $M(\vartheta_i; \xi)$ is the Fisher information matrix. The (v; s)th element of the Fisher information matrix can be written as

$$[M(\vartheta; \xi)]_{vs} = -\int \frac{\partial^2}{\partial \vartheta_v \partial \vartheta_s} \log(P_i(\beta))\xi(d\beta). \tag{7}$$

If the total number of items presented to the examinee is *N*, then the large sample variance estimate of $\hat{\mu}_{.5}$ is $\text{var}(\hat{\mu}_{.5}) / N$.

Ford et al. (1992) stated that c-optimal design for estimating a percentile consists of one difficulty level β_c if c is proportional to some functions of $z_c = \delta_i(\theta_i - \beta_c)$ for $-2.399 \le z_c \le 2.399$ for the logit function, or $-1.575 \le z_c \le 1.575$ for the probit function. In this particular case, to minimize the variance of $\hat{\theta}_i$, a design is considered c-optimal if it has the difficulty level, β_c, such that $P_i(\beta_c) = 0.5$.

D-optimal Design

The D-optimal design is used when one wishes to estimate both parameters δ_i and θ_i, the case in this study where δ_i is a slope parameter denoting the difference of the person's probabilities in correctly responding to easy and difficult items, and θ_i is a specified trait level. When done repeatedly for all θ levels in the examinee population, the result is an estimate of the entire person response function $P_i(\beta)$. The local D-optimality criterion guarantees (asymptotically) a minimum volume of the confidence ellipsoid (Karlin & Studden, 1966) for linear models and is used as a criterion for nonlinear models as well.

A design ξ^* is locally D-optimum if over all designs,

$$\det\{M(\vartheta_i;\xi^*)\} = \sup_{\xi\in D} \det\{M(\vartheta_i;\xi)\}. \qquad (8)$$

The D-optimal criterion actually prescribes the set of difficulty levels that maximize the determinant of Fisher's information matrix (cf. Atkinson and Donev, 1996). Due to the complexity of Fisher's information matrix, the analytical determination of these designs has proved difficult.

However, the D-optimal design has been derived for commonly used ogive models including the logit and probit of IRT applications. It has the two support points, β_1 and β_2, such that $P_i(\beta_1) = 0.083$ and $P_i(\beta_2) = 0.917$ for the logistic model (White, 1975) and $P_i(\beta_1) = 0.058$ and $P_i(\beta_2) = 0.942$ for the probit model (Abdelbasit & Plackett, 1983).

Sequential Design

The problem that an optimal design for nonlinear models depends on unknown parameters (the case with the person response models) can be resolved through sequential approximations. Sequential statistical procedures are those in which the next experiment is determined using the results from a previous trial. Up-and-down methods offer one way of sequentially administering test items. It is also a novel application using PRFs and may provide information useful when developing tests in educational and psychological domains.

Up-and-Down Designs

In psychometric studies, up-and-down designs are useful for estimating θ_i corresponding to a prespecified difficulty level Γ, $0 < \Gamma < 1$. They may also be used to estimate an entire person-response function, and they may do so much more efficiently than with D-optimal designs, as demonstrated below. It is important to realize at this point that the up-and-down designs do not rely on specific form of the models and that the parametric person response relationships are useful in evaluating or examining the performance of the design in educational adaptive testing situations.

The practical, defining characteristics of an up-and-down design are twofold, including (1) a finite set of possible items with difficulty levels, say $\Omega_B = \{\beta_1,...,\beta_K; \beta_1 < ... < \beta_K\}$, and (2) after an initial item is administered, the next item has either the same difficulty or one level higher or lower.

Let p_{lj} denote the probability of assigning difficulty level j given the last item was at difficulty l. Given that an examinee receives an item with difficulty β_k, denote the probability that the next items administered have difficulties β_{k-1}, β_k and β_{k+1} by $p_{k,k-1}$, $p_{k,k}$ and $p_{k,k+1}$, respectively, with $p_{k,k-1} + p_{k,k} + p_{k,k+1} = 1$ and $p_{1,-1} = p_{K,K+1} = 0$. The $p_{k,k-1}$, $p_{k,k}$ and

$p_{k,k+1}$ are transition probabilities. In a classical up-and-down design, difficulty levels cluster around the median difficulty level (ED50), as described below.

The Classical Up-and-Down Design

Lord (1970) applied the classical up-and-down method to adaptive testing for measurement purposes. Select the first item with a certain difficulty, i.e., set $B(1) = \beta_k$, for some $\beta_k \in \Omega_B$, where $B(1)$ is random or fixed [e.g., $B(1)$ might be the difficulty level thought to be closest to the target θ_Γ]. Then given $B(n) = \beta_k$, proceed sequentially as follows:

For $B(n) = \beta_k$, $k = 2, \ldots, K\text{-}1$,

$$B(n+1) = \begin{cases} \beta_{k-1}, & if \;\; Y(n) = 0 \\ \beta_{k+1}, & if \;\; Y(n) = 1 \end{cases}. \tag{9}$$

For $B(n) = \beta_1$,

$$B(n+1) = \begin{cases} \beta_1, & if \;\; Y(n) = 0 \\ \beta_2, & if \;\; Y(n) = 1 \end{cases}. \tag{10}$$

For $B(n) = \beta_K$,

$$B(n+1) = \begin{cases} \beta_{K-1}, & if \;\; Y(n) = 0 \\ \beta_K, & if \;\; Y(n) = 1 \end{cases}. \tag{11}$$

Then, defining $Q_i(\beta) = 1 - P_i(\beta)$, the procedure conforms to the following transition probability matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & \cdots & 0 \\ p_{21} & p_{22} & p_{23} & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & p_{K-1,K-2} & p_{K-1,K-1} & p_{K-1,K} \\ 0 & \cdots & 0 & p_{K,K-1} & p_{KK} \end{bmatrix} = \begin{bmatrix} Q_i(\beta_1) & P_i(\beta_1) & 0 & \cdots & 0 \\ Q_i(\beta_2) & 0 & P_i(\beta_2) & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & Q_i(\beta_{K-1}) & 0 & P_i(\beta_{K-1}) \\ 0 & \cdots & 0 & Q_i(\beta_K) & P_i(\beta_K) \end{bmatrix}. \tag{12}$$

Evaluating up-and-down designs, Lord (1970, 1971) concluded that they are more efficient for higher and lower trait level examinees. The classical up-and-down design targets the median difficulty level. For D-optimal designs, other difficulty levels must be targeted and for these, the Biased Coin Design (BCD) of Durham and Flournoy (1994) can be adopted. The BCD has a number of advantages in that it can target any arbitrary fractile, converges quickly, and has minimum variance among a large class of up-and-down designs (Bortot & Giovagnoli, 2005).

The Biased Coin Up-and-Down Design

The BCD is illustrated here. Let *h* be the probability that a biased coin comes up *head*. Fix *h* as a function of the odds of the correct response rate as follows (Durham & Flournoy,

1994):

$$h = \begin{cases} \dfrac{\Gamma}{1-\Gamma}, & 0<\Gamma \le 0.5 \\ \dfrac{1-\Gamma}{\Gamma}, & 0.5 \le \Gamma < 1.0 \end{cases}. \tag{13}$$

Again, select the first item with a certain difficulty, i.e., set $B(1) = \beta_k$, for some $\beta_k \in \Omega_B$,, where $B(1)$ is random or fixed. Then given $B(n) = \beta_k$, proceed sequentially as follows:

1. For $0 < \Gamma \le 0.5$: $B(n) = \beta_k$, $k = 2, \ldots, K\text{-}1$,

$$B(n+1) = \begin{cases} \beta_{k-1}, \text{ if } Y(n) = 0 \\ \beta_k, \text{ if } Y(n) = 1 \text{ and coin flip yields tails} \\ \beta_{k+1}, \text{ if } Y(n) = 1 \text{ and coin flip yields heads} \end{cases}. \tag{14}$$

For $B(n) = \beta_1$,

$$B(n+1) = \begin{cases} \beta_1, \text{ if } Y(n) = 0 \text{ or } \{Y(n) = 1 \text{ and coin flip yields tails}\} \\ \beta_2, \text{ if } Y(n) = 1 \text{ and coin flip yields heads} \end{cases}. \tag{15}$$

For $B(n) = \beta_K$,

$$B(n+1) = \begin{cases} \beta_{K-1}, \text{ if } Y(n) = 0 \\ \beta_K, \text{ if } Y(n) = 1 \end{cases}. \tag{16}$$

2. For $0.5 \le \Gamma < 1.0$: $B(n) = \beta_k$, $k = 2, \ldots, K\text{-}1$,

$$B(n+1) = \begin{cases} \beta_{k-1}, \text{ if } Y(n) = 0 \text{ and coin flip yields heads} \\ \beta_k, \text{ if } Y(n) = 0 \text{ and coin flip yields tails} \\ \beta_{k+1}, \text{ if } Y(n) = 1 \end{cases}. \tag{17}$$

For $B(n) = \beta_1$,

$$B(n+1) = \begin{cases} \beta_1, \text{ if } Y(n) = 0 \\ \beta_2, \text{ if } Y(n) = 1 \end{cases}. \tag{18}$$

For $B(n) = \beta_K$,

$$B(n+1) = \begin{cases} \beta_{K-1}, \text{ if } Y(n) = 0 \text{ and coin flip yields heads} \\ \beta_K, \text{ if } Y(n) = 1 \text{ or } \{Y(n) = 0 \text{ and coin flip yields tails}\} \end{cases}. \tag{19}$$

Continue sequentially until a stopping criterion is met.

In the special case that $\Gamma = 0.5$, $h = 1$ (the coin always comes up heads), the BCD reduces to the classical up-and-down design described above.

Defining $\bar{h} = 1 - h$, the BCD has transition probability matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & \cdots & \cdots & 0 \\ p_{21} & p_{22} & p_{23} & \cdots & \cdots & 0 \\ 0 & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & p_{K-1,K-2} & p_{K-1,K-1} & p_{K-1,K} \\ 0 & \cdots & \cdots & 0 & p_{K,K-1} & p_{KK} \end{bmatrix}$$

$$= \begin{bmatrix} Q_i(\beta_1) + \bar{h}P_i(\beta_1) & hP_i(\beta_1) & 0 & \cdots & \cdots & 0 \\ Q_i(\beta_2) & \bar{h}P_i(\beta_2) & hP_i(\beta_2) & \cdots & \cdots & 0 \\ 0 & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & Q_i(\beta_{K-1}) & \bar{h}P_i(\beta_{K-1}) & hP_i(\beta_{K-1}) \\ 0 & \cdots & \cdots & 0 & Q_i(\beta_K) & P_i(\beta_K) \end{bmatrix} \qquad (20)$$

for $0 < \Gamma \le 0.5$, whereas for $0.5 \le \Gamma < 1.0$, it is

$$P = \begin{bmatrix} Q_i(\beta_1) & P_i(\beta_1) & 0 & \cdots & \cdots & 0 \\ hQ_i(\beta_2) & \bar{h}Q_i(\beta_2) & P_i(\beta_2) & \cdots & \cdots & 0 \\ 0 & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & hQ_i(\beta_{K-1}) & \bar{h}Q_i(\beta_{K-1}) & P_i(\beta_{K-1}) \\ 0 & \cdots & \cdots & 0 & hQ_i(\beta_K) & \bar{h}Q_i(\beta_K) + P_i(\beta_K) \end{bmatrix}. \qquad (21)$$

As long as $P_i(\beta)$ is bounded away from 0 and 1 for all $\beta \in \Omega_B$, this matrix is regular, i.e., there exists some *n* for which all elements in the matrix $\mathbf{P}_n$ are positive. Because regular random walks converge exponentially fast to their stationary distributions, asymptotic results hold for moderately small sample sizes. This is useful, since many educational and psychological testing scenarios comprise small groups of examinees.

Asymptotic Distribution of Difficulty Levels Administered

Durham and Flournoy (1994) also described how Markov chain theory can be used in the biased coin up-and-down design. Here, let $\pi_k = \lim_{n\to\infty} P_i(B(n) = \beta_k)$ be the asymptotic (stationary) probability that an item of difficulty β_k will be selected. Let $N_k(n)$ denote the number of items (out of *n*) having difficulty level β_k; and, finally, $R_k(n)$ denotes the number of those answered correctly. Then, π_k is also the limiting proportion of items having difficulty level β_k, i.e., $\lim_{n\to\infty} N_k(n)/n$. Therefore, we call the set $\pi = \{\pi_k, k = 1, ..., K\}$ the *asymptotic difficulty distribution*. The asymptotic difficulty probabilities can be calculated from the transition probabilities as follows (Karlin & Taylor, 1975 p.107):

$$\pi_k = \prod_{j=2}^{k} \frac{p_{j-1,j}}{p_{j,j-1}}, \quad k = 2, \ldots, K,$$

$$\pi_1^{-1} = 1 + \sum_{k=2}^{K} \prod_{j=2}^{k} \frac{p_{j-1,j}}{p_{j,j-1}}. \tag{22}$$

Before initiating a test, it is advisable to explore expected outcomes under a variety of hypothetical response functions. Inserting prior estimates of the difficulty probabilities $\{P_i(\beta_k)\}$ into Equations 20 or 21 yields prior estimates of the transition probabilities and inserting these into Equation 23 yields prior estimates of the asymptotic difficulty distribution $\{\pi_k\}$. Thus, one can evaluate the large-sample performance of the experiment under a variety of possible scenarios, keeping in mind that reducing the interval between the possible difficulties will reduce the spread of the difficulty distribution.

It is necessary to make certain reasonable assumptions to obtain the desired properties of the difficulty scale parameter β, as mentioned above. Suppose that only item difficulty (β) levels are known. Additionally, given an examinee *i* with fixed θ_i, we assume that the PRF is a non-increasing function, i.e., the probability of endorsing an easier item correctly is higher than the probability of endorsing a more difficult item correctly so that $P_i(\beta_k) > P_i(\beta_{k+1})$. Given that the probability of correct response decreases with the increase in difficulty, the asymptotic difficulty distribution is unimodal with mode between θ_Γ and the θ levels adjacent to θ_Γ (Durham & Flournoy, 1994). Thus, difficulties are clustered around the θ having target difficulty level Γ.

To define the mode of the asymptotic difficulty distribution explicitly, let $M(\pi)$ denote the set of all difficulties in Ω_B having absolute maximum probabilities in the set $\{\pi_k\}$ and call $M(\pi)$ the modal set. Define the distribution π to be unimodal if there exists at least one integer *M* such that

$$\pi_k \geq \pi_{k-1} \text{ for all } k \leq M$$
$$\pi_{k+1} \geq \pi_k \text{ for all } k \geq M \tag{23}$$

The minimum value in the modal set, denoted β_M, is called the mode of the difficulty distribution. If the interval between difficulties is a constant Δ, then $|\beta_M - \theta_\Gamma| \leq \Delta/2$, that is, the most frequent difficulty is within $\pm\Delta/2$ of the target θ_Γ. In addition, the spread of difficulties is completely determined by the slope parameter δ_i of the response function model and by size of the intervals between difficulty levels.

Up-and-Down Designs for Approximating Optimal Designs

As noted earlier, locally non-sequential optimal designs administer items at some level that was assumed to be optimal. However, their performance relies on a fairly accurate prior specification of the unknown parameters. To circumvent the problem, sequential approximations are used. Up-and-down designs, targeting arbitrary fractiles (e.g., quantiles), do not require prior specification of the parameters and thus can be adopted to approximate optimal designs for parameter estimation of person response functions (Giovagnoli & Pintacuda, 1998). Mugno et al. (2004) used up-and-down procedures to estimate several percentiles simultaneously and found that the multiple-objective adaptive designs are more efficient and very robust against poor and/or biased prior information.

This study approximated both c-optimal and D-optimal designs. The former involves only

one design level for estimating $\hat{\theta}_i$, hence, only one up-and-down design is needed to approximate it. Actually, this approximation is the same as using the up-and-down design to approximate the median difficulty level. On the other hand, for the D-optimal design, two up-and-down designs were used with one targeting each optimal design level. Denote by ξ' an experiment consisting of two distinct sequential designs ξ_1' and ξ_2' with equal sample sizes. Then the information $M(\vartheta_i;\xi')$ for the whole experiment is

$$M(\vartheta_i;\xi') = \frac{1}{2}M(\vartheta_i;\xi_1') + \frac{1}{2}M(\vartheta_i;\xi_2'). \tag{24}$$

Examples

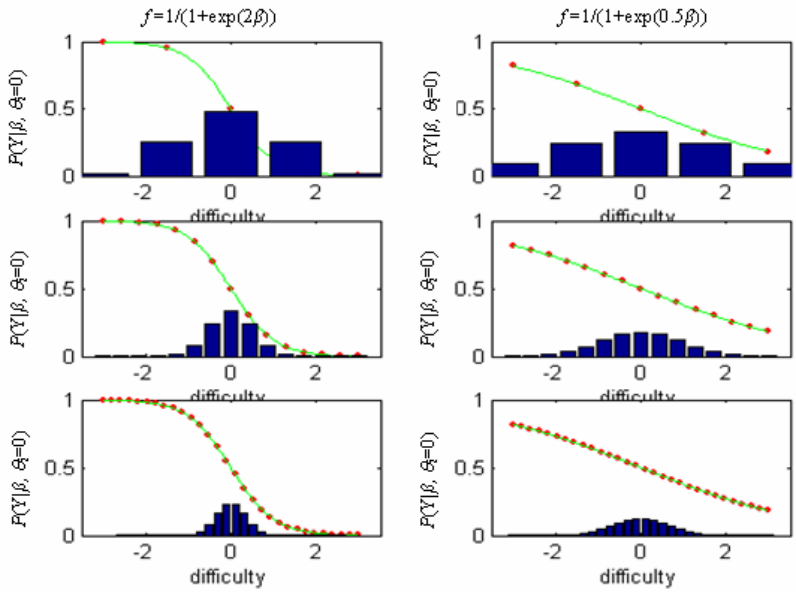
Examples are given in this section to illustrate the asymptotic difficulty distributions as well as the efficiency of up-and-down design approximating optimal designs.

Asymptotic Difficulty Distribution

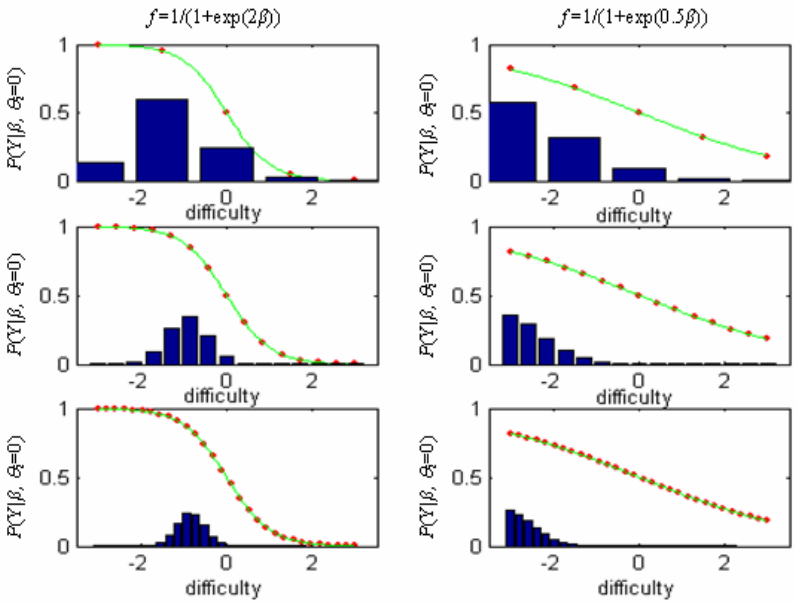
Asymptotic distributions of difficulty levels used in adaptive testing are obtained and plotted in Figures 2a to 2d for the classical up-and-down design, BCD targeting at the 17.6th quantile, BCD targeting at the 82.4th quantile, and up-and-down procedures approximating the D-optimal design, respectively. Three test situations were considered where an examinee with θ_i = 0 was given 5, 15, or 30 items. The item bank contained items with difficulty levels equally distant from β = -3 to 3. For example, with 5 items, the difficulty levels were β = -2.4, -1.2, 0, 1.2, and 2.4. In addition, the slope parameter δ_i is taken to be 2 or 0.5.

Figure 2. Asymptotic Distributions of Difficulty Under Three Different Situations When the Slope is 2 (Left Panels) or the Slope is 0.5 (Right Panels)

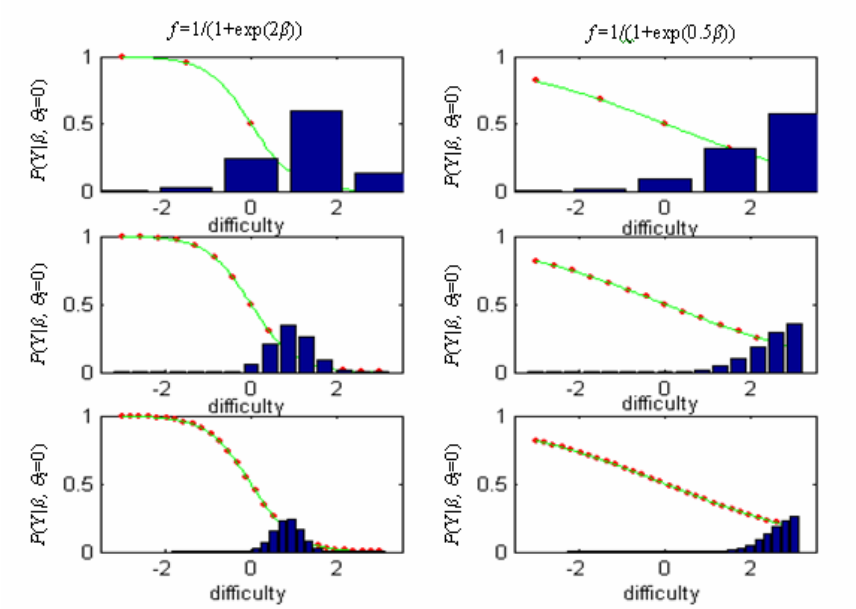
a. Items Having 5, 15 and 30 Peculiar Difficulties With the Classical Up-and-Down Design (Γ=.5)



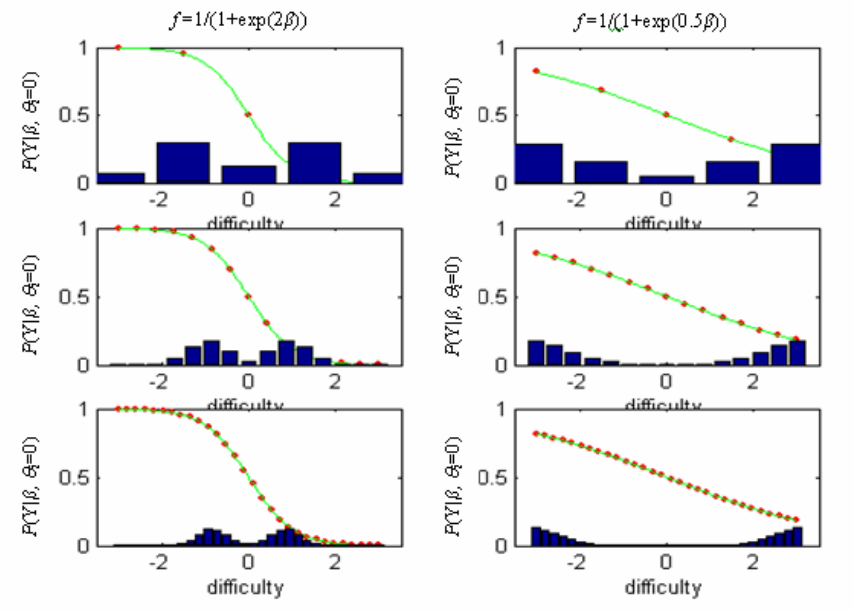
b. Items Having 5, 15 and 30 Distinct Difficulties With the Biased Coin Design (Γ=.176)



c. Items Having 5, 15 and 30 Distinct Difficulties
With the Biased Coin Design (Γ=.824)



d. Items Having 5, 15 and 30 Distinct difficulties
With Up-and-Down Approximating D-Optimal Design Points (Γ=.176 & Γ=.824)



Consistent with theory, the plots demonstrate three important features. First, both classical up-and-down design and BCD have asymptotic distributions with one mode. The mode is at $\theta_i = 0$ for the classical up-and-down design, and it shifts to left and right for the biased coin designs targeting the 17.6 and 82.4 quantals, respectively. Second, the up-and-down method approximated the two level D-optimal design; however, it resulted in a bimodal distribution which is a mixture of the distributions given in Figures 2b and 2c. Third, as the number of

distinct difficulty levels increased, the asymptotic difficulty distributions were increasingly smooth and targeted the quantal(s) more precisely. Fourth, the smaller the slope in the person-response function, the more widely spread the difficulty distributions were.

Up-and-Down Design Approximating Optimal Designs

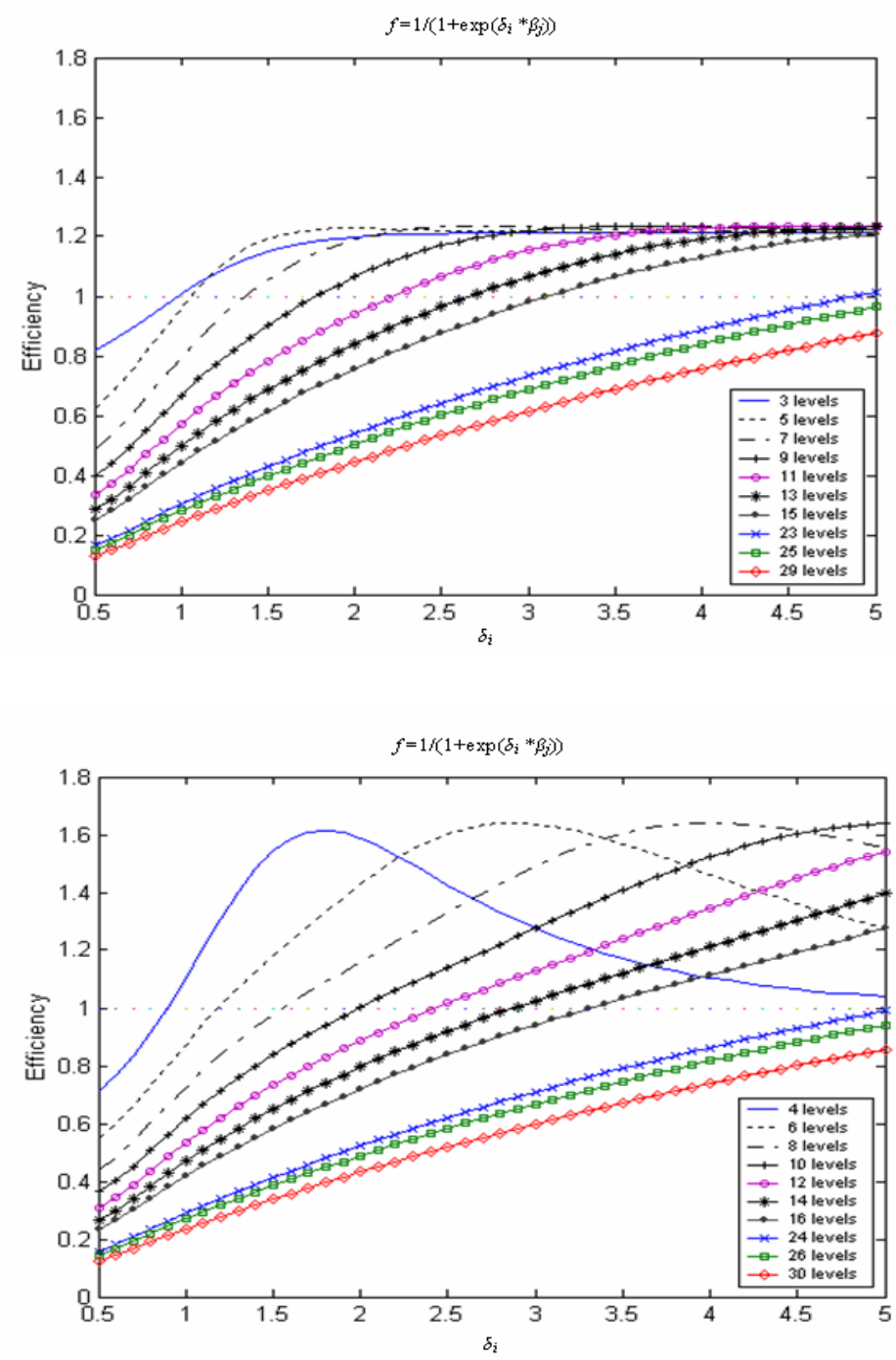
This section compares the up-and-down design approximating the c-optimal design with that approximating the D-optimal design. A criterion is to be defined for the comparison and it usually depends on different situations. When the main interest is on the entire person-response curve, a smaller determination of the variance-covariance matrix (det{M-1}) indicates relatively more efficiency of the design. On the other hand, when the main interest is on only $\hat{\theta}_i$, the median difficulty level variance ($\mathrm{Var}(\hat{\mu}_{.5})$) is used as the criterion for the comparison. Sometimes experimental data are used for purposes other than those considered in the experimental design. This section explores the effect of such use.

In the examples that follow, situations were considered with various distinct difficulty levels equally spaced from β = -3 to 3 and with different slopes ranging from δ = 0.5 to 5 for persons with θ_i = 0 or persons with θ_i two standard deviations above the mean, θ_i = 2.

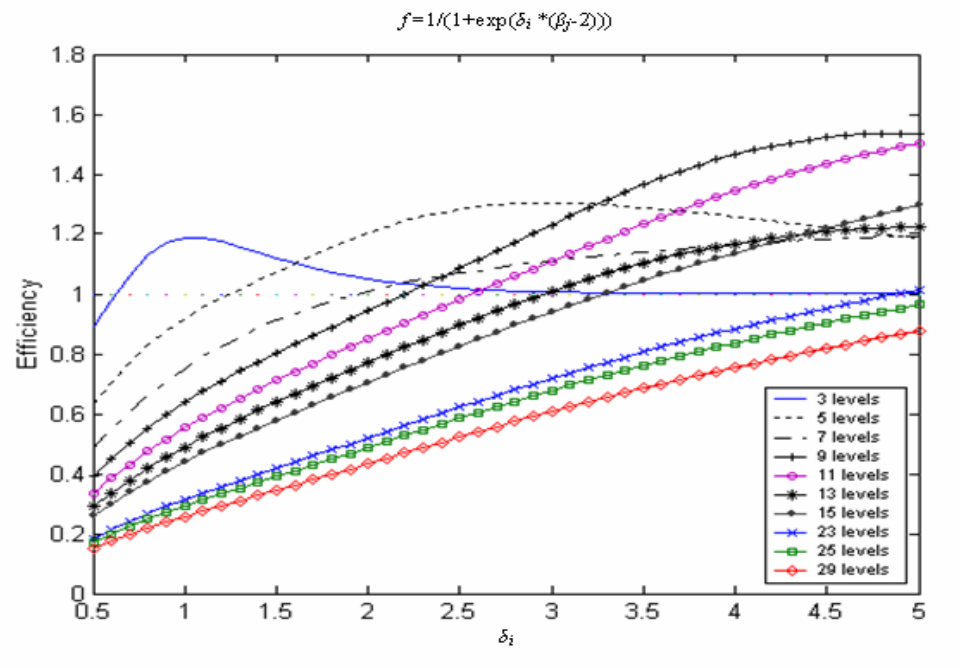
Whole curve efficiency (det{M-1}). As mentioned previously, when the main interest is in the entire curve, the determinant of the variance-covariance matrix is used to compare various optimal designs. The efficiency of the c-optimal approximation relative to the D-optimal approximation, which was obtained by taking the ratio of the det{M-1} for the D-optimal approximation to the det{M-1} for the c-optimal approximation, is displayed in Figure 3a for persons with θ = 0 and in Figure 3b for persons with θ = 2 . For ease of demonstration, the odd and even numbers of difficulty levels are separated into two plots. Any values above 1 indicate greater whole curve efficiency for the c-optimal approximation and values below 1 indicate greater whole curve efficiency for the D-optimal approximation. From asymptotic theory for exact designs all values were expected to be less than 1.0.

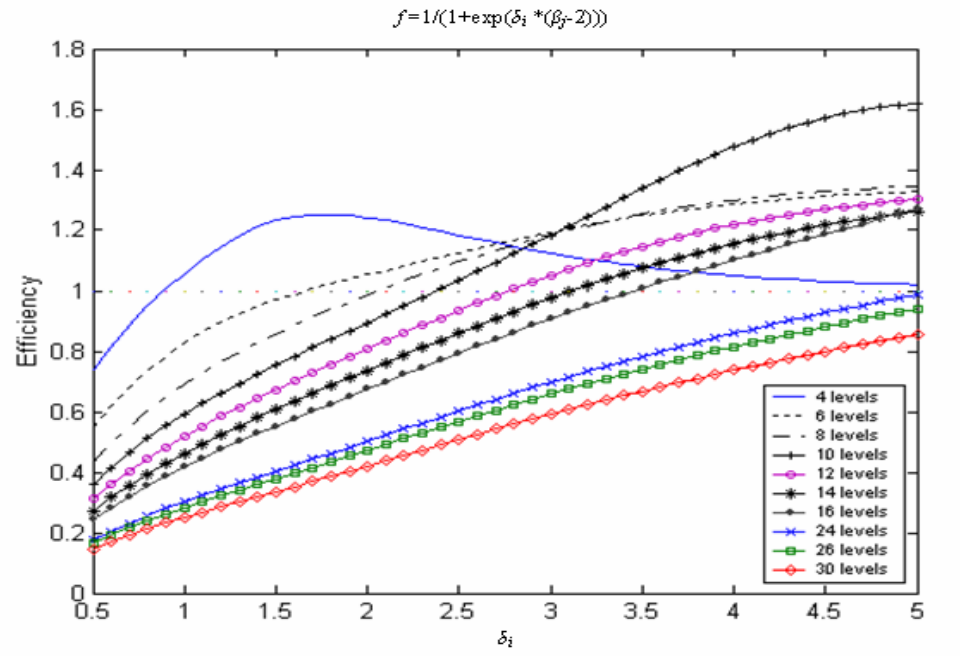
Figure 3. The Efficiency [det(M-1)] of the Up-and-Down Approximation for the c-Optimal Design Relative to the Up-and-Down Approximation for the D-Optimal Design

a. Items With a Different Number of Distinct Difficulty Levels are Administered to Examinees with $\theta_i = 0$



b. Items With a Different Number of Distinct Difficulty Levels are Administered to Examinees With $\theta_i = 2$





Generally, the two figures indicate that

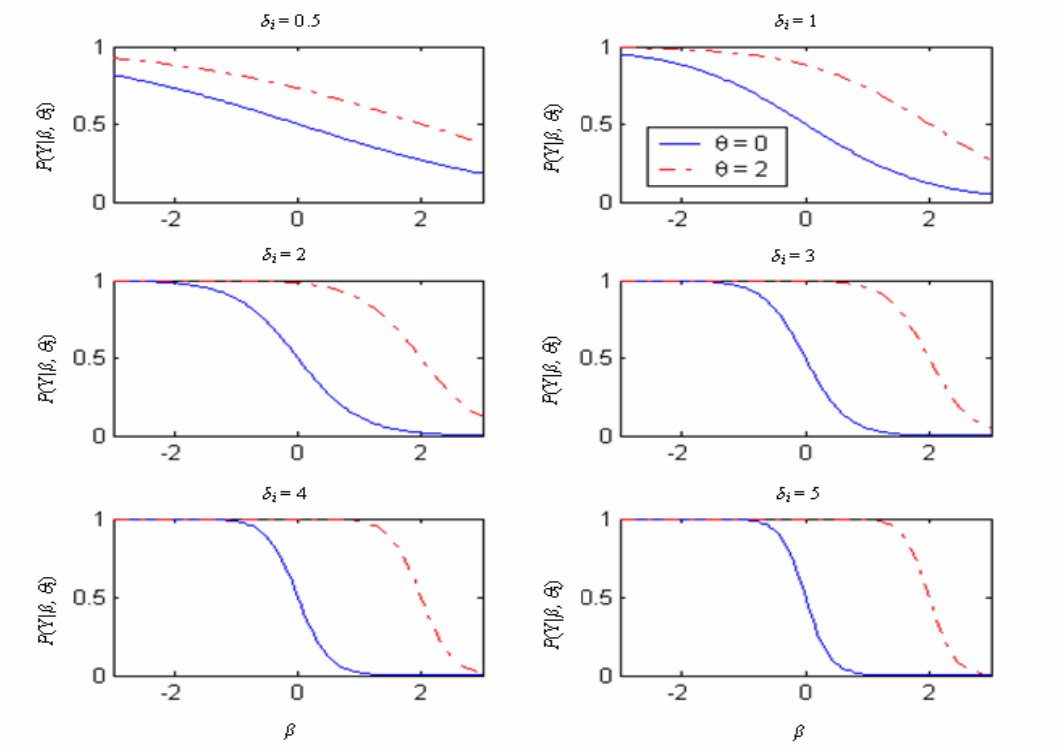
1. When the slope, δ, is small (smaller than 1), i.e., when the person's probabilities of correctly endorsing easy and difficult items are not much different, the D-optimal approximation is more efficient across the whole curve with different slopes, and therefore is preferred to the c-optimal approximation regardless of number of distinct difficulty levels.
2. The D-optimal approximation also is more efficient than the c-optimal approximation

when there are 23 or more distinct difficulty levels, regardless of how different are the person's probabilities of correctly responding to easy and difficult items.

3. For situations with small numbers of distinct difficulty levels and large slopes, δ, the c-optimal approximation is more likely to be preferred to the D-optimal approximation.
4. When comparing plots across the two θ levels, it is easy to see that with 15 or more distinct difficulty levels, the efficiency line patterns are generally the same. It is with small numbers of difficulty levels that efficiency line patterns differ. The difference appears between odd and even number of difficulty levels as well as between persons with an average θ and persons with a high θ.

To further understand the whole curve efficiency, Figure 4 displays the person-response curves with different slopes δ_i (δ_i varies from 0.5 to 5) for persons with $\theta = 0$ and 2. The solid lines are person-response curves for persons with $\theta = 0$ and the dotted lines are those for persons with $\theta = 2$. Comparing the solid lines across the six plots, it can be seen that as the slope δ_i becomes steeper, small numbers of difficulty levels fail to contain information to describe the whole curve. Therefore, the D-optimal approximation is less efficient than the c-optimal approximation. The dotted curves are the curves for persons with $\theta = 0$ shifted toward the left. Allowing for this shift, we see the same effect. That is, multiple difficulty levels in the steeply descending region of the person-response function are required in order to have efficiency (in the sense of D-optimality).

Figure 4. Person Response Functions With Different Slopes (δ_i = 0.5, 1, 2, 3, 4 and 5) and θ_i = 0 (Solid Curve) or θ_i = 2 (Dashed Curve)



It can hence be understood that when obtaining the whole curve efficiency displayed in Figure 3b, various distinct difficulty levels ranging from -3 to 3 were again used for the shifted curves. Therefore, as the curves are steeper, small numbers of distinct difficulty levels

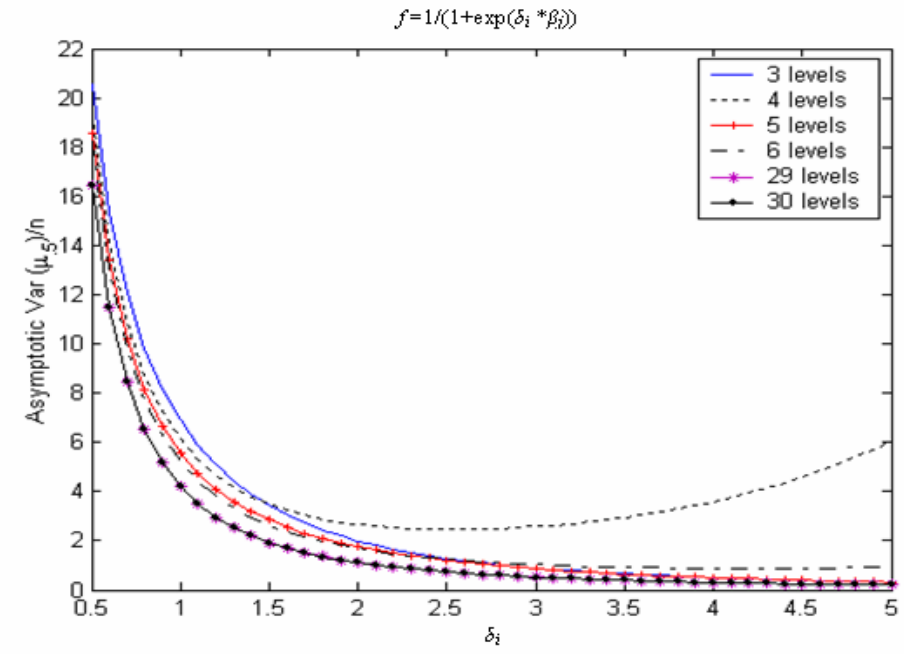
are less likely to gather enough information to estimate the whole curve. Hence PRFs with small differences for a person's probabilities in correctly endorsing easy and difficult items (presumably $\delta_i \leq 1$) are recommended for D-optimal approximation when the interest is on estimating the entire curve.

Median Difficulty Level Variance ($\mathrm{Var}(\hat{\mu}_{.5})$). Sometimes the main interest is only in estimating the median difficulty level for a person, $\hat{\mu}_{.5}$ (or $\hat{\theta}_i$ denoted previously), instead of the entire person response curve. Then, the variance or the standard deviation of this particular percentile estimate is a useful measure of design quality. The design with smaller variance $\mathrm{Var}(\hat{\mu}_{.5})$ is said to be more efficient in estimating the median difficulty level. According to this definition, it is believed that the c-optimal approximation should be more efficient in estimating this percentile.

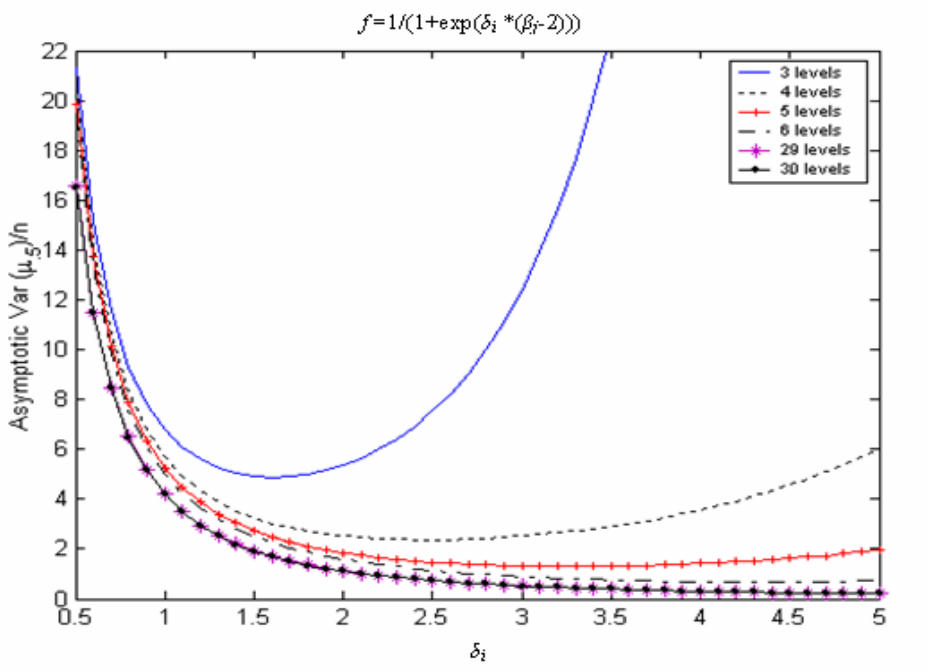
Median difficulty level variances are plotted for the c-optimal approximation in Figure 5a and Figure 5b as a function of the slope δ_i for persons with $\theta = 0$ and 2, respectively. Generally, the variance gets smaller with a higher number of distinct difficulty levels and with larger slope. Almost all variances are smaller than 22, except for the extreme situation in which item banks have only 3 distinct difficulty levels, $\theta = 2$, and the slope of the person-response curve is steeper than $\delta_i = 3.5$. This point can be further illustrated using Figure 4. That is, using 3 difficulty levels equally spaced from $\beta = -3$ to 3 (i.e., $\beta = -3, 0, 3$), the median difficulty level is estimated less precisely as the slope gets steeper, for less information can be gathered from only a small portion of the shifted curve.

Figure 5. Asymptotic $\mathrm{Var}(\hat{\mu}_{.5})$ for the Up-and-Down Approximation to the c-Optimal Design (Γ=.5)

a. $\theta_i = 0$



b. $\theta_i = 2$



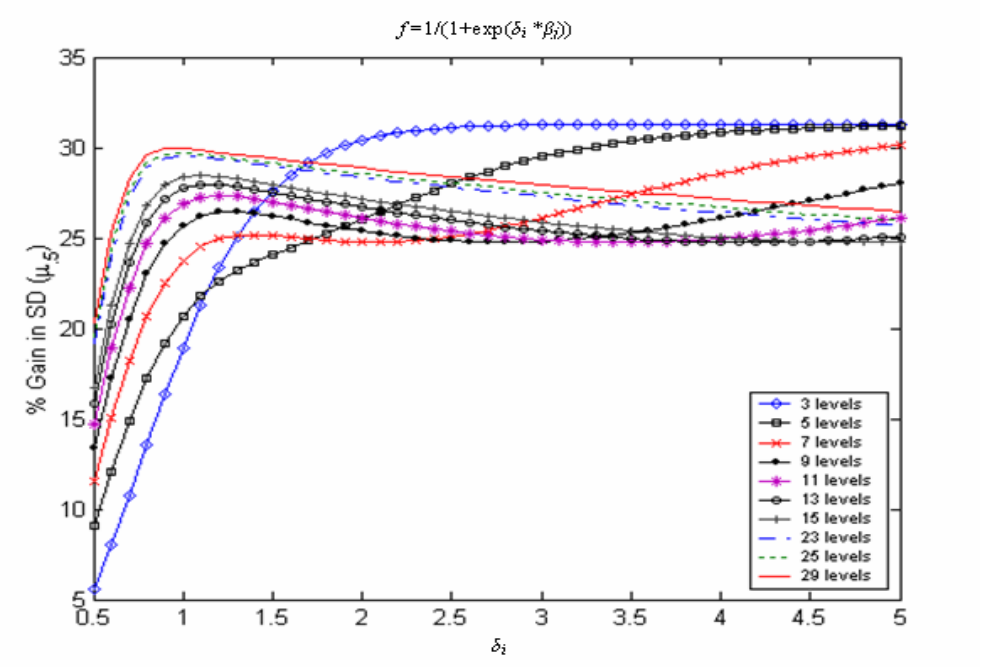
To compare the c- and D- optimal design approximations, we considered the gain in the standard deviation of $\hat{\mu}_{.5}$ defined by

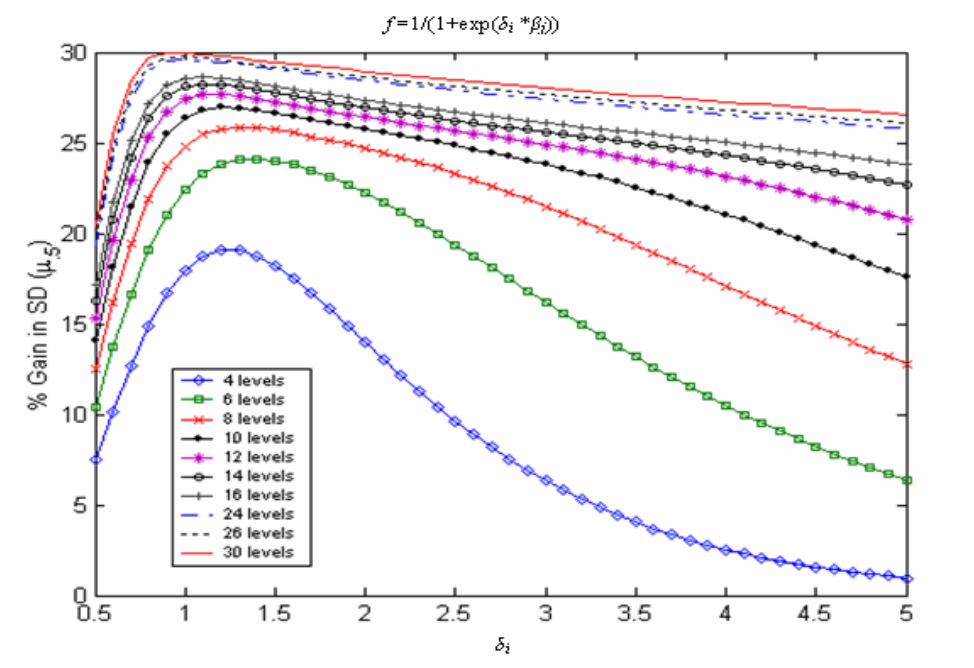
$$SD(\hat{\mu}_{.5}) = \frac{SD_{D-opt.} - SD_{c-opt.}}{SD_{c-opt.}}. \qquad (25)$$

Figures 6a and 6b plot the percent gain in $SD(\hat{\mu}_{.5})$ for persons with $\theta_i = 0$ and $\theta_i = 2$, respectively. It is suggested from the figures that (1) the percent gains in $SD(\hat{\mu}_{.5})$ were almost all positive, indicating that the D-optimal approximation was less efficient than the c-optimal approximating in estimating the median difficulty level alone; (2) the average gain was about 22% for $\theta_i = 0$ and 15% for $\theta_i = 2$; (3) with more distinct difficulty levels, the c-optimal approximation was more desirable than the D-optimal approximation; and (4) the pattern for small numbers of distinct difficulty levels varied across the plots. However, when there were 13 or more difficulty levels and the slope was larger than 2.5, the pattern of gain in $SD(\hat{\mu}_{.5})$ was similar for persons with different θ levels.

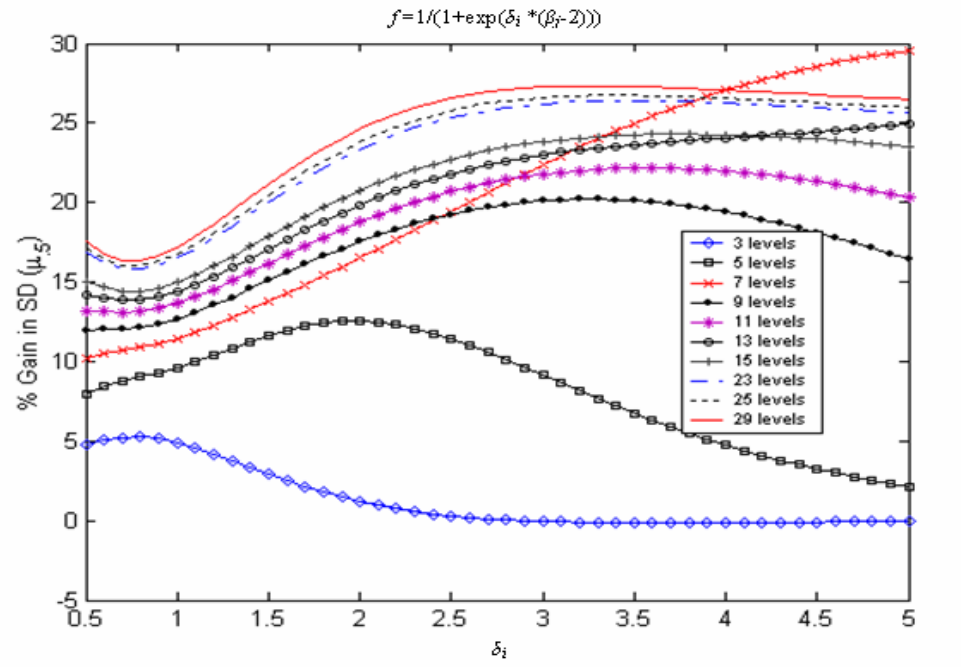
Figure 6. Gain in $SD(\hat{\mu}_{.5})$ by Using the Up-and-Down Approximation to the D-Optimal Design Instead of the Up-and-Down Approximation to the c-Optimal Design

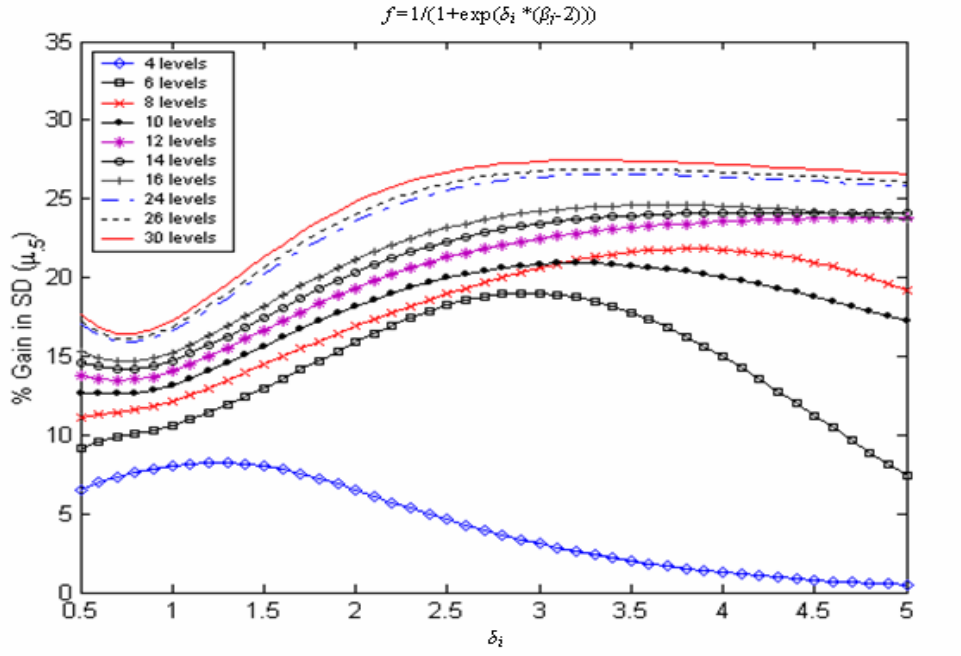
a. $\theta_i = 0$





b. $\theta_i = 2$





Conclusions

For optimal item selection procedure in adaptive testing, optimal designs can be constructed to target the median difficulty level (c-optimality) or the entire person-response curve (D-optimality). The current study used sequential up-and-down designs from the person response perspective to approximate the c- or D-optimal design. It is concluded that both approximation procedures performed well in various adaptive testing situations. Comparing the two procedures, the up-and-down design approximating the c-optimal design was more efficient when the main focus was on estimating the median difficulty level,

whereas the up-and-down design approximating the D-optimal design was more efficient when the main focus was on estimating the whole person-response curve. In effect, the latter worked better when there were no less than 23 distinct difficulty levels. With small numbers of difficulty levels, it is yet unclear as to which procedure performed consistently better. Therefore, tests are recommended to be constructed with large numbers of distinct difficulty levels when D-optimality is of interest.

This study illustrated an item selection procedure from the person response perspective, which uses sequential up-and-down designs to target an arbitrary percentile of interest or the entire person response functions. This new approach leaves open a series of questions that need to be addressed, such as: how the up-and-down design compares with the conventional maximum information algorithm for item selection; how the person response perspective compares with the item response perspective; and how practical issues such as starting point, stopping rule, item exposure , etc., affect the estimation of person trait levels.

References

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

Atkinson, A. C. & Donev, A.N. (1996). *Optimum experimental designs.* Oxford University Press: Oxford

Berger, M. P. F. (1992) 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. & Mathijssen, E. (1997). Optimal test designs for polytomously scored items. *British Journal of Mathematical and Statistical Psychology, 50,* 127-141.

Buyske, S. G. (1998). *Optimal design for item calibration in computerized adaptive testing.* Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Bortot, P., & Giovagnoli, A. (2005). Up-and-down experiments of first and second order. *Journal of Statistical Planning and Inference, 134,* 236-253.

Dixon, W. J., & Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association, 43,* 109-126.

Durham, S. D., & Flournoy, N. (1994). Random walks for quantile estimation. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics* (pp. 467-476). New York: Springer-Verlag.

Fedorov, V. V. & Hackl, P. (1997). *Model-oriented design of experiments.* New York: Springer-Verlag.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd Ltd.

Ford, I., Torsney, B., & Wu, C. F. J. (1992). The use of a canonical form in the construction of locally optional designs for non-linear problems. *Journal of the Royal Statistical Society, Series B, 54,* 569-583.

Giovagnoli, A., & Pintacuda, N. (1998). Markovian experiments that approximate optimal designs for quantal response curves. *Metron,* 56, 77-96.

Karlin, S., & Studden, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics.* New York: John Wiley & Sons.

Karlin, S., & Taylor, H. M. (1975). *A first course in stochastic processes* (2nd ed.). New York: Academic Press.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper and Row.

Lord, F. M. (1971). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association, 66,* 707-711.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31,* 19-26.

Mugno, R., Zhu, W., & Rosenberger, W. F. (2004). Adaptive urn designs for estimating several percentiles of a dose-response curve. *Statistics in Medicine, 23*, 2137-2150.

Pukelsheim, F. (1993). *Optimal design of experiments.* New York: Wiley.

Rosenberger, W. F., Flournoy, N., & Durham, S. D. (1997). Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *Journal of Statistical Planning and Inference, 60,* 69-76.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191-208.

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

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

Thissen, D., & Mislevy, R. J. (2000). Testing Algorithms. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, R. Mislevy, L. Steinberg & D. Thissen (Eds.), *Computerized Adaptive Testing: A Primer* (pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New York: Academic Press.

von Bèkèsy, G. (1947). A new audiometer. *Acta Otolaryngology, 35*, 411-422.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (Eds.). *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

White, L. V. (1973). An extension of the general equivalence theorem to nonlinear models. *Biometrika, 60,* 345-348.

White, L. V. (1975). *The optimal design of experiments for estimation in nonlinear models*. Ph.D. thesis, University of London.