Graduate
Management
Admission
Council®

Item Selection With Biased-Coin Up-and-Down Designs

Yanyan Sheng
Southern Illinois University

Zhaohui Sheng
Western Illinois University

Presented at the Item Selection Paper Session, June 2, 2009

Abstract

Targeting any arbitrary percentile, the biased-coin up-and-down design is theoretically appealing and can provide an efficient alternative to the current maximum Fisher information method for item selection in adaptive testing. This paper illustrates the use of the design with the one-parameter item response model and further evaluates its utility by comparing it with the conventional method in a few simulated conditions. Results from simulation studies indicate that the biased-coin up-and-down design is flexible in targeting any difficulty level, and that it outperforms the conventional item selection method in certain circumstances.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Sheng, Y. and Sheng, Z. (2009). Item selection with biased-coin up-and-down designs. In D. J. Weiss (Ed.), *Proceedings of the 2009 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. Email: ysheng@siu.edu

Item Selection With Biased-Coin Up-and-Down Designs

As computerized adaptive testing (CAT) is becoming increasingly popular, it has been implemented in various testing programs in and outside the United States. A basic ingredient in CAT is the item selection procedure that sequentially selects and administers items based on a person's responses to the previously administered items. For decades, the maximum Fisher information (MFI; Lord, 1977) criterion has been widely used as the conventional method for item selection in CAT (e.g., Thissen & Mislevy, 2000) because it is conceptually simple and theoretically efficient (e.g., Kingsbury & Zara, 1989).

However, item selection by Fisher's information is not ideal. This method selects items to match the provisional person trait (θ) estimate, so that each person has about a .5 probability of endorsing the next item correctly. Hence, with an optimal item bank, about 50 percent of the items are expected to be answered correctly. However, in some testing situations, it might be more desirable to target a different probability . For example, when young children are being measured, it might be appropriate to administer an instrument so that a larger proportion of items can be answered correctly to keep them highly motivated in the content area. Alternatively, in some testing situations, a mastery level test might be desired so that test developers are able to specify a certain proportion of items that examinees can endorse correctly. Fisher information item selection is, consequently, limited in certain CAT applications. In addition, the reliance on the prior θ estimates in the selection of the subsequent items to be administered increases the need for accurate estimation procedures. Nonetheless, none of the currently used estimation procedures performs well in all testing situations. For example, maximum likelihood estimation (MLE; Lord, 1977) does not work well in short tests (e.g., Hambleton & Swaminathan, 1985) in that it tends to give rise to multiple local maxima for tests with less than 20 items (Lord, 1980). On the other hand, Bayesian estimation tends to shrink toward the prior mean, and hence rely on correct specifications of the prior distribution, as misspecfication can lead to large estimation bias (e.g., Gorin, Dodd, Fitzpatrick, & Shih, 2005).

Up-and-down designs are sequential designs and have been widely used in bioassay applications. They initially received attention in the 1940s (Anderson, McCarthy, & Tukey, 1946) for quantile estimation. Dixon and Mood (1948) focused on such designs specifically for estimating the 50th percentile of the dose response function; they were later called the classical up-and-down designs (Lord, 1970). Up-and-down designs have been studied by many in the biostatistical community, including Wetherill (1963), Dixon (1965), Wetherill and Glazebrook (1986), Tsutakawa (1967, 1980), Storer (1989), Flournoy (1990), and Durham and Flournoy (1993), among others, before Durham and Flournoy (1994) proposed and popularized the biased-coin up-and-down design (BCD), which is a simple sequential design that requires smallest possible sample sizes without loss of estimation accuracy. The BCD has a number of advantages in that it can target any arbitrary percentile (not just the 50th percentile), converges quickly, and has minimum variance among a large class of up-and-down designs (Bortet & Giovagnoli, 2005).

CAT shares many similarities with the adaptive testing in bioassay. Specifically, while one controls item parameters such as α_j, β_j, and/or γ_j to estimate θ in CAT, the bioassayist controls θ to estimate the value of β, which is defined, in biostatistics, as the dosage level at which a certain percentage of the treated subjects is toxicized. Given this, it is reasonable to believe that the item selection algorithm based on the BCD, which does not rely on an accurate θ estimate in every step of CAT administrations, provides an efficient alternative to, while being more flexible than, the conventional item selection method.

In the CAT literature, early efforts have been made to apply the classical up-and-down design to adaptive testing in educational and psychological measurement (e.g., Lord, 1970; 1971). Due to its lack of improved measurement accuracy over a linear paper-and-pencil test and the fact that more importance had been attached to the information function in item response theory (IRT), attention has been focused on other item selection algorithms. The BCD was recently introduced and applied to CAT from a person response perspective (Sheng, Flournoy, & Osterlind, 2007). Although it has shown promise in that context, it is not clear whether the BCD provides utility in IRT-based CAT. The purpose of this study was to illustrate the use of the BCD in CAT from the item response perspective and to further evaluate its utility by comparing it with the conventional MFI algorithm.

Item Selection With the BCD

The practical, defining characteristics of an up-and-down design are twofold, including (1) a finite set of possible item characteristic levels that can be arranged in order, i.e., $\Omega_X = \{\xi_1, ..., \xi_K; \xi_1 < ... < \xi_K\}$, and (2) after an initial item is administered, the next item has either the same level on the item characteristic(s) under study or one level higher or lower. Hence, depending on the number of characteristics that are used to distinguish between items in the bank, ξ_k can be a scalar or a vector.

The BCD Item Selection Algorithm

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):

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

Select the first item with a certain characteristic level, i.e., set $X(1) = \xi_k$, for some $\xi_k \in \Omega_X$, where $X(1)$ is random or fixed. Then given $X(\ell) = \xi_k$, the BCD proceeds sequentially as follows:

1. For $0 < \Gamma \leq 0.5$, if $X(\ell) = \xi_k$, k = 2,…, K– 1,

$$X(\ell+1) = \begin{cases} \xi_{k-1}, & \text{if } Y(\ell) = 0 \\ \xi_k, & \text{if } Y(\ell) = 1 \text{ and coin flip yields tails} \\ \xi_{k+1}, & \text{if } Y(\ell) = 1 \text{ and coin flip yields heads} \end{cases}; \tag{2}$$

   if $X(\ell) = \xi_1$,

$$X(\ell+1) = \begin{cases} \xi_1, & \text{if } Y(\ell) = 0 \text{ or } \{Y(\ell) = 1 \text{ and coin flip yields tails}\} \\ \xi_2, & \text{if } Y(\ell) = 1 \text{ and coin flip yields heads} \end{cases}; \tag{3}$$

   if $X(\ell) = \xi_K$,

$$X(\ell+1) = \begin{cases} \xi_{K-1}, & \text{if } Y(\ell) = 0 \\ \xi_K, & \text{if } Y(\ell) = 1 \end{cases}. \tag{4}$$

2. For $0.5 \leq \Gamma < 1.0$, if $X(\ell) = \xi_k$, k = 2,…, K – 1,

$$X(\ell+1) = \begin{cases} \xi_{k-1}, & \text{if } Y(\ell) = 0 \text{ and coin flip yields heads} \\ \xi_k, & \text{if } Y(\ell) = 0 \text{ and coin flip yields tails} \\ \xi_{k+1}, & \text{if } Y(\ell) = 1 \end{cases}; \tag{5}$$

   if $X(\ell) = \xi_1$,

$$X(\ell+1) = \begin{cases} \xi_1, & if\ \ Y(\ell) = 0 \\ \xi_2, & if\ \ Y(\ell) = 1 \end{cases}; \tag{6}$$

   if $X(\ell) = \xi_K$,

$$X(\ell+1) = \begin{cases} \xi_{K-1}, & \text{if } Y(\ell) = 0 \text{ and coin flip yields heads} \\ \xi_K, & \text{if } Y(\ell) = 1 \text{ or } \{Y(\ell) = 0 \text{ and coin flip yields tails}\} \end{cases}. \tag{7}$$

This procedure continues sequentially until a stopping criterion is met.

In the special case where $\Gamma = 0.5$, $h = 1$ (the coin always comes up heads), the BCD simplifies to the classical up-and-down design (see Lord, 1970 for a detailed illustration).

IRT Model

It is important to note here that the BCD does not rely on a model of a specific form and that the parametric item response relationships are useful in evaluating or examining the performance of the design in educational CAT situations. For ease of implementation of the BCD from an item response perspective in CAT, this study focused on the one-parameter IRT model, which takes the form

$$P_j(\theta) = P(Y_j = 1 | \theta) = \frac{1}{1+e^{-(\theta-\beta_j)}}, \quad (8)$$

where β_j denotes the *j*th item difficulty level. This implies that item difficulty is the only item characteristic that differentiates between different items, namely, $\xi_j = \beta_j$.

Simulation Studies

Method

To investigate the utility of the BCD for item selection in CAT, two Monte Carlo simulation studies were conducted in which either a fixed- or a random- stopping rule was employed. With the fixed-stopping rule, the item bank was fixed to have items with 100 different difficulty levels and the CAT stopped when *k* (*k* = 5, 10, 30, 100) items had been administered. With the random-stopping rule, the item bank had items with *n* (*n* = 10, 30, 50, 100) different difficulty levels and the CAT stopped when the test information reached 3.5 (i.e., the standard error of measurement reached approximately .535). In either case, item difficulties were randomly generated from a $U(-2,2)$ distribution, and CAT responses were simulated for persons whose actual θ levels were 0 (the average), −1 (1 standard deviation below the average), and −2 (2 standard deviations below the average) using the IRT model specified in Equation 8. Further, subsequent items were selected based on each of the following four procedures: (1) the MFI, (2) the BCD with $\Gamma = 0.2$, (3) the BCD with $\Gamma = 0.5$, and (4) the BCD with $\Gamma = 0.8$, with the three BCD procedures targeting at the 80th, the 50th and the 20th percentiles in item difficulty levels, respectively. Each CAT simulation began θ estimation with an initial value of 0 and used MLE.

With either stopping rule, 10,000 replications were conducted to reduce sampling error. The accuracy of person parameter estimates was evaluated using the mean square error (*MSE*) and *bias*. Let $\hat{\theta}_r$ denote the estimated person θ parameter in the *r*th replication (*r* = 1, …, *R*). The *MSE* is defined as

$$MSE = \frac{\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2}{R}, \quad (9)$$

and the *bias* is defined as

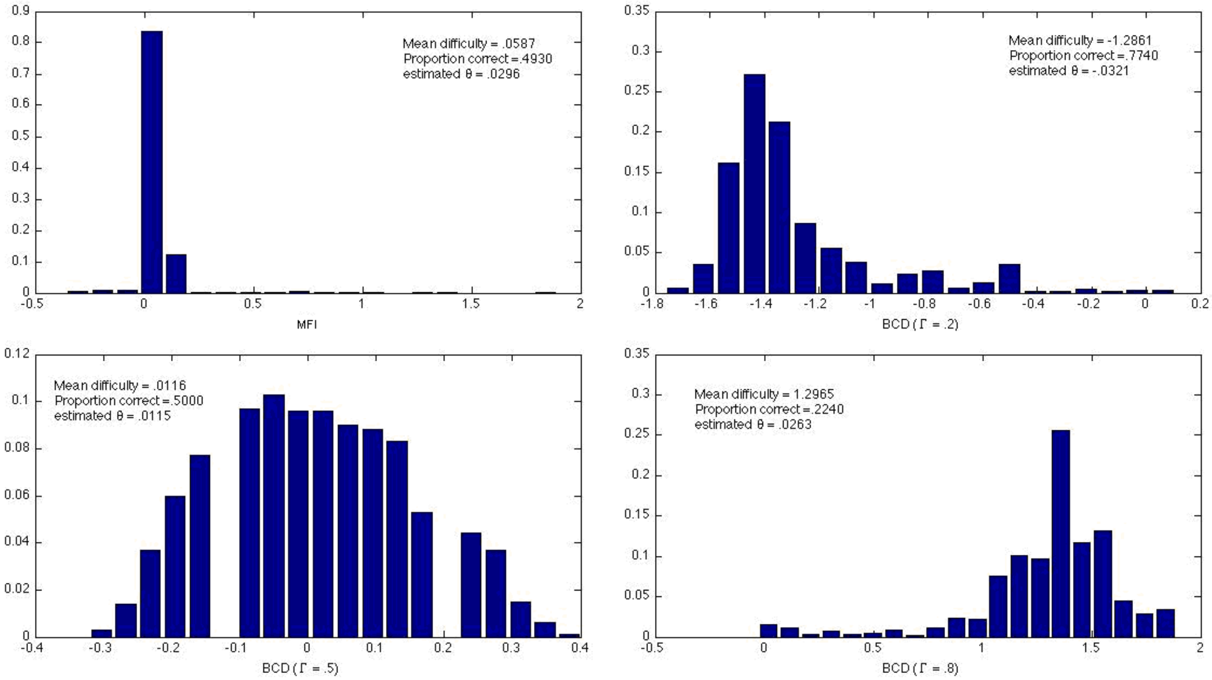
$$bias = \frac{\sum_{r=1}^{R}(\hat{\theta}_r - \theta)}{R}. \quad (10)$$

Results

Before the *MSE* and *bias* results are summarized, the four item selection procedures were compared using a single simulation with the fixed-stopping rule, one replication at *θ* = 0, 100 difficulty levels in the item bank (*n*), and 1,000 items administered (*k*) was used here simply to illustrate theoretical properties of items selected by each procedure.

The difficulty levels for items administered using the four procedures were obtained and their density plots are displayed in Figure 1, from which it can be seen that 80% of time, the MFI selected items with a medium difficulty level (i.e., β = 0). On the other hand, the BCD with $\Gamma = 0.5$, though targeting at the medium difficulty level on average, differed from the MFI in that the items selected had difficulties ranging from –.3 to .4. This potentially allows the person to be exposed to a variety of items. The BCD with $\Gamma = 0.2$ selected items with difficulties that ranged from –1.8 and 0 and centered around –1.3, whereas the BCD with $\Gamma = 0.5$ selected relatively more difficult items, with difficulties ranging from 0 to 2 and an average of 1.3. Given the item difficulty distributions, it is not surprising to observe that proportions of correct responses using the four selection procedures were 49.3%, 77.4%, 50%, and 22.4%, respectively. Hence, the flexibility of the BCD in targeting any percentile in the difficulty levels has been illustrated. The four procedures resulted in similar MLE estimates of θ, which were .0296, –.0321, .0115, and .0263, respectively. These estimates were fairly close to the true θ value of 0 in this simulation, with the BCD with $\Gamma = 0.5$ relatively closer. Additional results based on 10,000 replications summarized below support this finding.

Figure 1. Empirical Density Plots of the Difficulty Values for Items Selected Using the Four Procedures in a CAT Simulation Where $\beta_j \sim U(-2,2)$, $\theta = 0$, *n* = 100, *k* = 1000



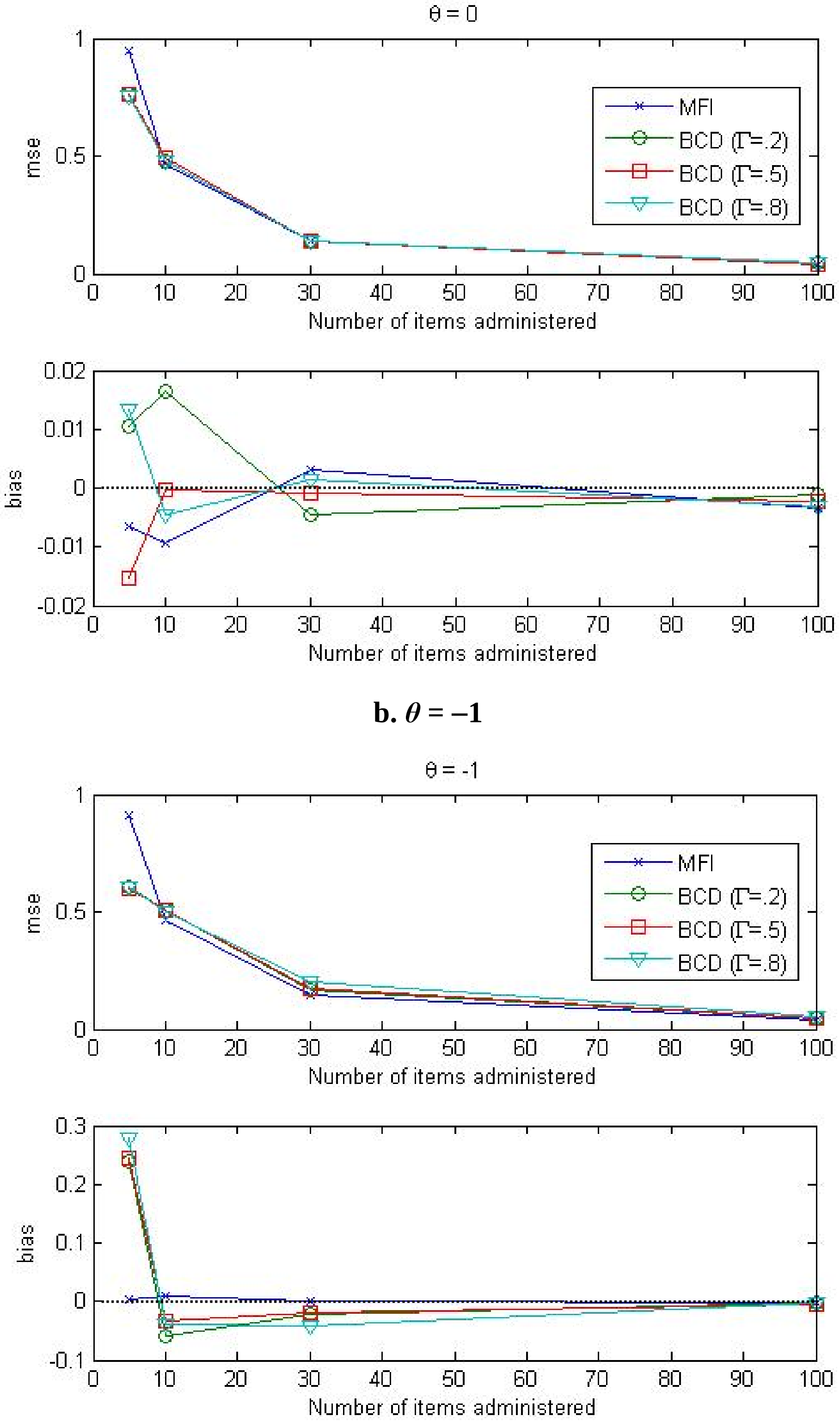
For simulations with the fixed-stopping rule, the *MSE* and *bias* results are plotted in Figure 2 for θ = 0, θ = –1, and θ = –2, respectively. The following observations can be made from the plots:

1. When θ = 0, the MFI had a relatively larger *MSE* but smaller *bias* than the three BCD procedures with *k* = 5. As *k* increased, the four procedures were close in their *MSE* values, although the MFI had a larger bias. Among the three BCD procedures, the BCD with $\Gamma = 0.5$ seemed to perform relatively better for *k* > 5.

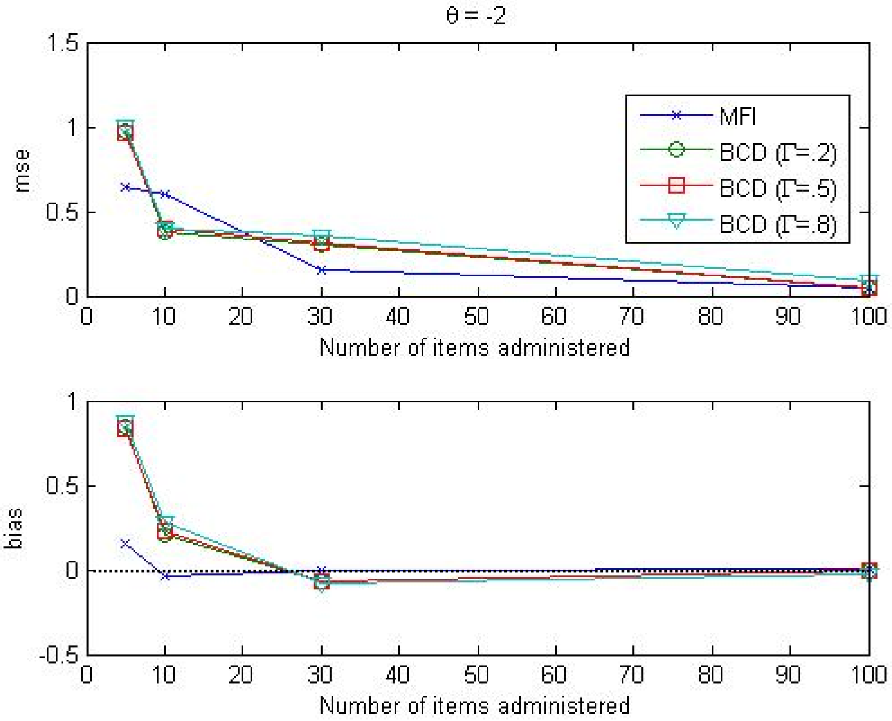
Figure 2. Parameter Recovery With the Fixed-Stopping Rule

a. *θ* = 0

b. *θ* = –1



c. $\theta = -2$



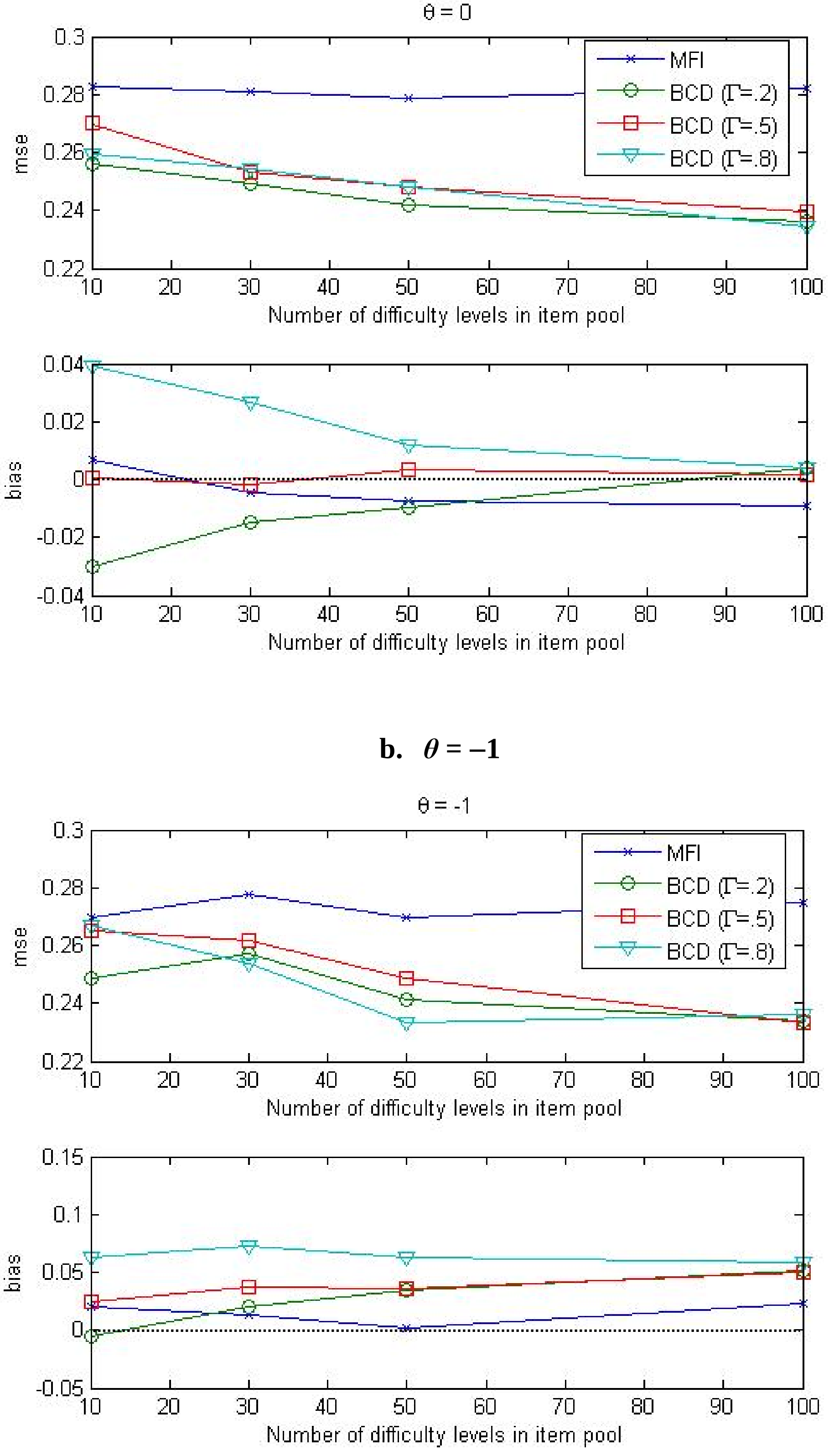
2. When $\theta = -1$, the MFI, having consistently the smallest *bias* for different *k* values considered, had a smaller *MSE* than the three BCD procedures for $k > 5$. The three BCD procedures performed similarly with respect to the *MSE*.
3. When $\theta = -2$, the MFI had relatively smaller *bias* than the three BCD procedures except for $k = 100$, where the BCD with $\Gamma = 0.5$ had the smallest *bias*. In addition, except for $k = 10$, the MFI had relatively smaller *MSE*. Among the three BCD procedures, the BCD with $\Gamma = 0.8$ had consistently larger *MSE* and *bias* values.

Hence, the results suggest that BCD did not show much advantage in the accuracy of estimating θ over the MFI using the fixed-stopping rule, particularly when θ was –2. It is noted that $\theta = -2$ was at the limit of the difficulty levels in the item bank as the difficulties were generated from $\mathrm{U}(-2, 2)$ in the simulations. Therefore, the BCD is not recommended for estimating person θ levels at or close to the limit of the difficulty levels in the item bank. When the actual θ level is at or close to the center location of the distribution of item difficulties, the BCD, especial the BCD with $\Gamma = 0.5$ is preferred to the MFI when the number of items administered is small.

For simulations with the random-stopping rule, the *MSE* and *bias* results are plotted in Figure 3 for $\theta = 0$, $\theta = -1$, and $\theta = -2$, respectively. From the plots, we can observe the following:

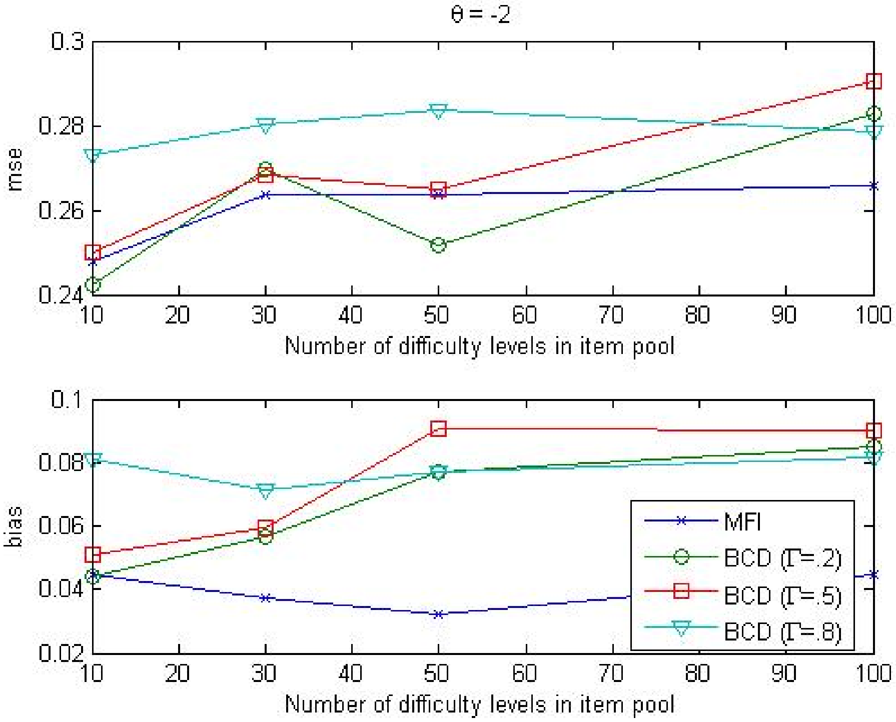
Figure 3. Parameter Recovery With the Random-Stopping Rule

a. *θ* = 0



b. *θ* = –1

c. *θ* = –2



1. When $\theta = 0$, the MFI consistently had a larger *MSE*, though not necessarily a larger *bias*, than the three BCD procedures, among which the BCD with $\Gamma = 0.2$ had relatively smaller *MSE* in all the situations considered. Further, the BCD with $\Gamma = 0.5$ consistently resulted in the smallest *bias*.

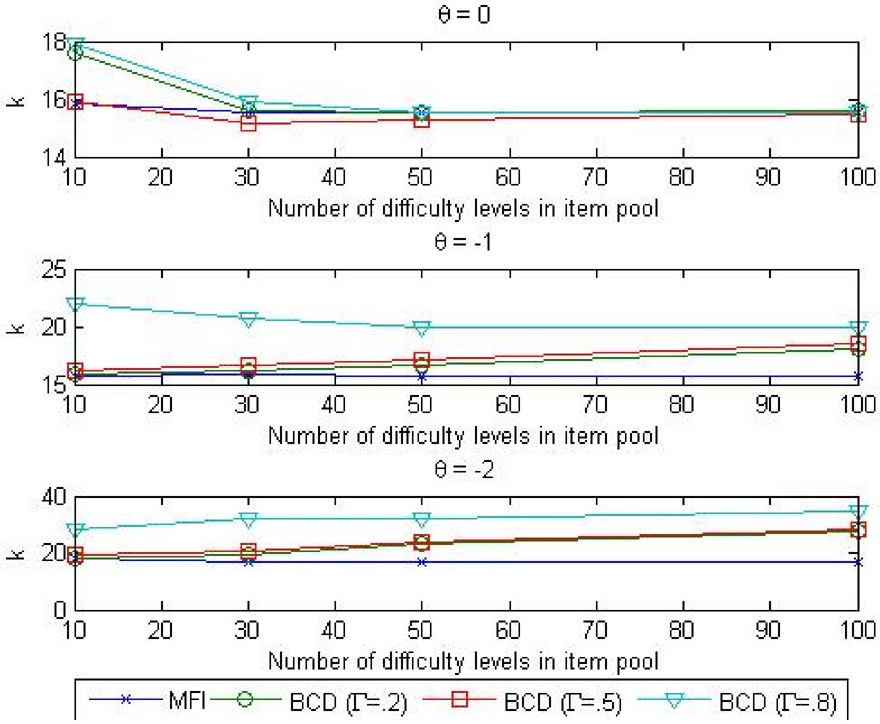
2. When $\theta = -1$, the MFI, with relatively a smaller *bias* for $n > 10$, had again consistently larger *MSE* than the three BCD procedures. Among the three BCD procedures, the BCD with $\Gamma = 0.2$ had relatively a smaller *bias*.

3. When $\theta = -2$, the MFI had consistently a smaller *bias* than the three BCD procedures. Further, when $n = 30$ and $n = 100$, the MFI had the smallest *MSE*, whereas when $n = 10$ and $n = 50$, the BCD with $\Gamma = 0.2$ had the smallest *MSE*. A close examination of the *MSE* values indicates that the three BCD procedures did not differ much in estimating *θ*, except that the BCD with $\Gamma = 0.8$ had a slightly larger *MSE* when $n < 100$.

Consequently, the three BCD procedures under study have shown advantages over the traditional MFI in estimating *θ* using the random-stopping rule. In particular, when *θ* was not at or close to the limit of the item difficulty levels in the item bank (such as –2 in the simulations), the BCD tended to result in a smaller *MSE*, though not a smaller *bias* compared with the MFI. It has to be noted that such advantages may come with a price, as the BCD tends to administer more items than the MFI using the random-stopping rule. This is understandable as the MFI always selects items that maximize item information.

Examination of the number of items administered (*k*) using each selection procedure (Figure 4) reveals that when θ was 0, MFI had *k* similar to or even larger than BCD (particularly BCD with $\Gamma = 0.5$); whereas when θ moved away from 0, BCD required more items than MFI as *n* increased. Among the three BCD procedures, the BCD with $\Gamma = 0.8$ consistently had a larger *k* and the other two BCD procedures were close for $\theta \neq 0$, with the BCD with $\Gamma = 0.2$ having a slightly smaller *k*. Therefore, considering the efficiency of the CAT, the BCD with $\Gamma = 0.5$ is recommended when the person's actual θ level is around 0.0. When the θ moves away from 0.0, the BCD with $\Gamma = 0.2$ may be adopted for item banks consisting of items with small number of different difficulty levels (such as *n* = 10 in the illustration).

Figure 4. Number of Items Administered Using Each Item Selection Procedure With the Random-Stopping Rule



Discussion

Proposing the biased-coin up-and-down design as an alternative item selection algorithm in CAT, this study illustrated the procedure from the item response perspective, i.e., using an item response function. Since the BCD requires the arrangement of item characteristic levels, the one-parameter IRT model was used for ease of implementation. Results from the simulation studies suggest that BCD is flexible enough to target any arbitrary percentile. Hence, it can be used for situations for which it is desired that a certain proportion of items is expected to be answered correctly. With respect to the accuracy in estimating the person's θ level, BCD performed equivalent to or even better than the conventional MFI algorithm when the actual person θ was not at or close to the limit of the item difficulties in the item bank. Hence, the results suggest that BCD is more flexible and should provide an efficient alternative to the conventional MFI method.

During the study, it was found that the BCD depends on the starting difficulty level, as subsequent items selected have a higher or lower difficulty level. It is, therefore, not very efficient when the item bank consists of items with a large number of different difficulty levels, as is seen from the simulation results with the random-stopping rule. Additional studies are needed to improve the BCD to make it more efficient. Moreover, as mentioned previously, this study only considered the simplest case where items differed only in their difficulty levels; they might also be different in their discrimination or guessing parameters. However, complexity arises in such circumstances with respect to selecting items using the BCD as each item characteristic level is a vector of two or three values. Future studies are needed to develop an up-and-down algorithm that works well with the two- or three-parameter IRT models.

References

Anderson, T. W., McCarthy, P. J., & Tukey, J. W. (1946). “Staircase” method of sensitivity testing. *Naval Ordinance Report.* Princeton: Statistical Research Group, 46-65.

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. (1965). The up-and-down method for small samples. *Journal of the American Statistical Association, 60,* 967-978.

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

Durham, S. D., & Flournoy, N. (1993). Convergence results for an adaptive ordinal urn design. *Journal of the Theory of Probability and its Applications, 37,* 14-17.

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

Flournoy, N. (1990). Adaptive designs in clinical trials. *Proceedings of the Biopharmaceutical Section of the American Statistical Association.* Alexandria, VA: American Statistical Association.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29,* 433-456.

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

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Psychological Measurement, 2,* 359-375.

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.

Lord, F. M. (1977). A board-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

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

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.*

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics, 45,* 925-937.

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* (*2nd ed*) (pp. 101-133). Hillsdale, NJ: Erlbaum.

Tsutakawa, R. K. (1967). Random walk design in bio-assay. *Journal of the American Statistical Association, 62,* 842-856.

Tsutakawa, R. K. (1980). Selection of dose levels for estimating a percent point of a quantal response curve. *Applied Statistics, 29,* 25-33.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society B, 25,* 1-48.

Wetherill, G. B., & Glazebrook, K. D. (1986). Sequential estimation of points on quantal response curves. *Sequential Methods in Statistics.* New York: Chapman & Hall.