A New Approach to Simulation Studies in Computerized Adaptive Testing

by

Shu-Ying Chen

National Ping-Tung Teachers College
Taiwan, R. O. C.

Paper presented at the 2001 Annual Meeting of the
American Educational Research Association, Seattle, WA. USA.

Abstract

Precision of trait estimation and determination of item exposure rates are always important considerations in conducting CATs. To date, simulation techniques have been used in research concerning these two considerations. Even though conducting simulation studies has some advantages over using real human subjects, the simulation results are not as accurate as analytically obtained results. To improve the accuracy of the results, and to avoid time consuming simulation work, more research on analytical solutions to problems is needed. The purpose of this study was to investigate the precision of CAT trait estimation and to determine CAT item exposure rates using an analytical approach (referred to as Tree). The results thus obtained were compared to those obtained using a simulation study, where three different levels of replications, 100, 500, and 1,000(referred to as S100, S500, and S1K respectively) were implemented.

Based on this study, the differences among S100, S500, S1K, and Tree decreased as test length increased. For test length as long as 20 items, the differences among S500, S1K, and Tree were negligible. In other words, a simulation study with 500 replications can provide results as accurate as the analytical approach for a 20-item test.

A New Approach to Simulation Studies in Computerized Adaptive Testing

Shu-Ying Chen
National Ping-Tung Teachers College
Taiwan, R.O. C.

Introduction

To date, the efficiency and precision afforded by CATs has typically been studied using computer simulation techniques (e.g., Veerkamp & Berger, 1997; Wang, & Vispoel, 1998). By conducting simulation studies, CAT components (i.e., item characteristics, item pool sizes, test lengths, item selection rules, test termination rules, etc.) can be controlled and manipulated easily, so the effects of factors of interest may be assessed more efficiently than by using real human subjects.

Even though conducting simulation studies has some advantages over using real human subjects, the simulation results are not as accurate as analytically obtained results. Furthermore, the simulation work is time consuming. To improve the accuracy of the results, and to avoid time consuming simulation work, research on analytical solutions to problems is needed. The purpose of this study was to investigate the precision of CAT trait estimation and to determine CAT item exposure rates using an analytical approach. The results thus obtained were compared to those obtained using a simulation study.

Theoretical Framework of an Analytical Approach

The analytical approach that was used in this study is illustrated with the following example. Consider a four-item CAT administered from a pool of ten items in which each item is defined by the dichotomous three-parameter logistic (3PL) item response model and content specifications as well.

The tree diagram in Figure 1 illustrates all possible four-item CATs that can be administered when no item exposure control is implemented. Each rectangle in the tree diagram is numbered according to the item number it represents, and each row of the diagram corresponds to a particular stage of the test (i.e., 1st, 2nd, 3rd, or 4th). The letter in parentheses within a rectangle represents the probability of the corresponding item being administered at the corresponding stage of the CAT. These probabilities are defined in Table 1. Without considerations of item exposure control, Item 5 is always administered at the first stage of the CAT in this case; hence, the probability of Item 5 being administered at the first stage must equal one. During the second stage, Item 6 is administered to an examinee only if Item 5 is answered correctly; therefore, $P(5)$ is the probability that Item 6 is administered to examinees at the second stage. On the other hand, if Item 5 is answered incorrectly in the first stage, then Item 4 is administered to an examinee at the second stage, with probability 1-$P(5)$. Note that the sum of the probabilities of the two items that can be administered at the second stage (i.e., $P(5)$ and 1-$P(5)$) is equal to one. During the third stage, Item 7 is administered only if Items 5 and 6 are both answered correctly.

Thus, the probability that Item 7 is administered at the third stage is *P*(5)×*P*(6), and the probability that Item 4 is administered at the third stage is *P*(5)×[1-*P*(6)]. Similarly, the probability that Item 6 is administered at the third stage is [1-*P*(5)]×*P*(4), and the probability that Item 3 administered at the third stage is [1-*P*(5)]×[1-*P*(4)]. Note that the four probabilities at the third stage also sum to one. Based on the same logic, the probabilities for the items administered to examinees at the fourth stage are obtained and sum to one.

Based on the probability structure defined by Table 1, the precision of trait estimation can be determined at any stage of the CAT, for any given true trait level. For example, assuming a true trait level $\theta = 3$, two possible trait estimates can be obtained after Item 5 has been administered: one if the item is answered correctly, and the other if it is answered incorrectly. The probability of answering Item 5 correctly (or incorrectly) is also known given $\theta = 3$. Therefore, a discrete probability distribution can be formed and the mean and variance of this distribution can be determined exactly without any approximation. Similarly, at any other stage of the CAT, a discrete probability distribution can be determined along with corresponding mean and variance. Thus, the precision of trait estimation for any given true trait level can be determined analytically at any stage of the CAT.

The item exposure rate of any item in the pool can also be determined using the probability structure defined by Table 1, simply by summing the probabilities associated with that item across all stages of the CAT. Table 2 shows these tabulations for each of the items in the CAT item pool. Thus, the item exposure rate of each item can also be determined analytically.

The results of this simple example can be readily extended to any item pool. They can also be applied under different conditions of initial trait estimates, item selection rules, or trait estimators. The utility of the analytical approach may be limited, however, in cases of long tests, where the tree diagram is big. For test length as long as 20 items, the number of all possible 20-item CATs is $2^{19} = 524{,}288$. Thus, it would be time consuming to obtain results for long tests using the analytical approach. For the 20-item test, it took a PC (Pentium III, 650Mhz, 128MB RAM) around 15 minutes to run for the analytical approach while only a couple of minutes for a simulation study with 1,000 replications.

Even though the analytical approach is time consuming for long tests, the accuracy of its results is undoubted while the accuracy of the simulation results is always unknown no matter how many replications are used, although the more the replications, the more accurate the results. Based on the results obtained from the analytical approach, the purpose of this study was to evaluate the precision of trait estimation and accuracy of item exposure rates obtained from a simulation study, where three different levels of replications, 100, 500, and 1000 were implemented.

Method

The procedure of the analytical approach has been described as above. The procedure of the simulation study is described in details in this section. For both approaches, the following specifications for the CAT components were applied.

Item response model. The three-parameter logistic (3PL) item response model, which is commonly used in operational CATs, was applied in the current study. This model defines the probability of a correct response, given a trait level (θ), as follows:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}, \qquad (1)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo chance level or pseudo guessing parameter.

Item pool structure. The item pool consisted of 360 ACT-Math items with item parameters calibrated using the 3PL item response model with data obtained from the years of 1993, 1994, and 1995. Descriptive statistics of the item parameters are shown in Table 3.

Initialization. Because no prior information about examinees' trait levels could be obtained before testing, the initial trait estimate was assumed to be zero (i.e., $\hat{\theta}_0 = 0$).

Trait estimation. Because of the problems associated with MLE, especially at the early stages of a CAT, EAP estimation with a θ~N(0,1) prior distribution was used in this study. The EAP estimate can be approximated using Gauss-Hermite quadrature (Stroud & Sechrest, 1966) as follows:

$$\hat{\theta} \equiv E(\theta \mid \mathbf{U}) = \frac{\sum_{k=1}^{q} X_k L(X_k) W(X_k)}{\sum_{k=1}^{q} L(X_k) W(X_k)} \tag{2}$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with X_k, $L(X_k)$ is the likelihood function conditioned at X_k, and $\mathbf{U} = (u_1, \ u_2, \ . \ . \ . \ , \ u_n)$ is a response vector.

Test length and termination rule. The maximum test length was set at 20 items. Data were collected at each unitary increment of test length from i = 1 to 20, inclusive. The comparison of these two approaches, however, was carried out for test length equal to 10, 15, and 20 items only.

Item selection. Maximization of item information was the criterion used for item selection at each stage. The item information is defined as follows:

$$I_j(\theta) = \left[\frac{\partial P_j(\theta)}{\partial \theta}\right]^2 \Big/ P_j(\theta) Q_j(\theta) = \frac{\left[P_j^{'}(\theta)\right]^2}{P_j(\theta) Q_j(\theta)} = \frac{2.89 a^2 (1-c)}{\left[c + e^{1.7a(\theta-b)}\right]\left[1 + e^{-1.7a(\theta-b)}\right]^2} \tag{3}$$

where $P_j(\theta)$ is the item response function for Item j and $Q_j(\theta) = 1 - P_j(\theta)$.

True Trait levels. The precision of CAT trait estimation and item exposure rates were studied at each of seven different true trait levels: θ_0 = -3, -2, -1, 0, 1, 2, and 3. These trait levels cover most of the range of trait levels observed in practice, and they are commonly applied in research (e.g., Veerkamp & Berger, 1997).

Simulation Study

The procedure of the simulation study with 1,000 replications is described as follows. Similar procedures are applied for 100 and 500 replications. At each trait level, 1,000 replications (i.e., simulees) were used for the CAT simulation. Each simulee-item response was generated based on the comparison of $P_j(\theta)$ from the 3PL (Equation 1) to a random

$U(0,1)$ deviate. If $P_j(\theta)$ was greater than or equal to the value of the random uniform deviate, then the simulee-item score was 1; otherwise, the simulee-item score was 0. The 1000 x 360 simulee-item response matrix was fixed at each true trait level studied to eliminate the effect of randomness of responses.

Given an EAP trait estimate based on n items having been administered so far ($\hat{\theta}_n$; $\hat{\theta}_0 = 0$ according to the initialization assumption stated above), the (n + 1)th item was selected such that $I_j\left(\hat{\theta}_n\right)$ had the maximum value among all of the items in the pool.

After an item was administered, the 1000 x 360 simulee-item response matrix described earlier was used to determine whether an examinee answered the item correctly. Then, a new EAP was obtained to identify the next appropriate item for the examinee. The process continued until a 20-item test had been administered.

Evaluation Criteria

Bias, standard error and item exposure rates were used to evaluate the results obtained from both the analytical approach and the simulation study. The procedure to find these evaluation criteria for the analytical approach has been described as above. The procedure for the simulation study is described as follows.

For each combination of test length (i.e., n = 1, 2, 3, . . . , 20) and true trait level (i.e., θ_o= -3, -2, -1, 0, 1, 2, 3), bias (BIAS) and standard error (SE) were calculated across 1,000 replications as follows:.

$$BIAS(n,\theta_o) = \frac{1}{1{,}000}\sum_{k=1}^{1{,}000}\left(\hat{\theta}_{n,k} - \theta_o\right) \tag{4}$$

$$SE(n,\theta_o) = \sqrt{\frac{1}{1{,}000}\sum_{k=1}^{1{,}000}\left(\hat{\theta}_{n,k} - \bar{\hat{\theta}}_n\right)^2},\ \bar{\hat{\theta}}_n = \frac{1}{1{,}000}\sum_{k=1}^{1{,}000}\hat{\theta}_{n,k} \tag{5}$$

At each true trait level, item exposure rate of each item was calculated. The item exposure rate of item i, is defined as

$$r_i = \frac{m_i}{p} \quad , \quad i = 1, 2, 3, \ldots n \tag{6}$$

where m_i is the number of times item i is presented to the simulees, p is the total number of simulees and n is the size of an item pool. To compare the item exposure rates obtained from both approaches, chi-square statistic was calculated as follows:

$$\chi^2 = \sum_{i=1}^{360}\frac{\left(r_{i,s1k} - r_{i,tree}\right)^2}{r_{i,tree}} \tag{7}$$

where $r_{i,s1k}$ is the item exposure rate of Item i obtained from the simulation study with 1,000 replications and $r_{i,tree}$ is the item exposure rate of Item i obtained from the analytical approach.

Results

SEs

Figure 2 summarizes the standard error results obtained from the analytical approach and the simulation study. In general, among the three levels of replication of the simulation study (S100, S500, and S1K), the SEs obtained from S1K were the closest to those obtained from the analytical approach (Tree). Furthermore, the longer the test, the closer the SEs. For test length as long as 20 items, they were not distinguishable. The relationship between S500 and Tree was similar to that observed for S1K. Compared to S100, the SEs obtained from S500 were closer to those obtained from Tree. Also, the longer the test, the closer the SEs. The differences were very slim when test length was equal to 20 items. That is, the differences among S500, S1K, and Tree were negligible for a 20-itesm test. Compared to the results observed above, the SEs obtained from S100 were quite different from those obtained from Tree. Except for the true trait level 2, SEs tended to be overestimated at negative true trait levels while underestimated at the other true trait levels when S100 was implemented. However, the difference decreased as test length increased. The differences were still noticeable when test length as long as 20 items.

Bias

Figure 3 summarizes the bias results obtained from the analytical approach and the simulation study. The relationships among the four approaches with respect to bias seemed not as distinguishable as those observed for SEs. In general, the bias obtained from S1K was the closest to that obtained from Tree and the longer the test, the closer the bias. The bias obtained from S1K and Tree were not distinguishable when test length was equal to 20 items. Compared to S100, the bias obtained from S500 was closer to that obtained from Tree. Also, the longer the test, the closer the bias. The differences were also very slim when test length as long as 20 items. Compared to SEs, the bias obtained from S100 was not quite different from that obtained from Tree. The differences were not noticeable when test length as long as 20 items. That is, with respect to bias, the differences among S100, S500, S1K, and Tree were negligible for a 20-item test.

Item Exposure Rate

Table 4 shows the Chi-square statistic for the simulation study with three different levels of replications. The results were consistent with those observed for SE and Bias. Among the three levels of replications, the item exposure rates obtained from S1K were the closest to those obtained from Tree, and thus associated with the smallest Chi-square values. Compared to S100, S500 had much smaller chi-square values and performed much more similarly to Tree for a 20-item test.

Conclusions & Discussion

Based on the results described above, it’s clear that the differences among S100, S500, S1K, and Tree decreased as test length increased. For test length as long as 20 items, the differences among S500, S1K, and Tree were negligible. In other words, a simulation study with 500 replications can provide results as accurate as the analytical approach for a 20-item test. Thus, to decide the number of replications needed for a simulation study, we may have to consider test length as an important factor. For test length longer than 20 items, less than 500 replications may be sufficient to provide accurate results. More research is needed to confirm this conclusion.

Table 1. Probability of each item administered at each stage

1st stage

Item 5: $a = 1.0$

2nd stage

Item 6: $b = a \times P(5)$
Item 4: $c = a \times [1 - P(5)]$

3rd stage

Item 7: $d = b \times P(6)$
Item 4: $e = b \times [1 - P(6)]$
Item 6: $f = c \times P(4)$
Item 3: $g = c \times [1 - P(4)]$

4th stage

Item 8: $h = d \times P(7)$
Item 4: $i = d \times [1 - P(7)]$
Item 7: $j = e \times P(4)$
Item 3: $k = e \times [1 - P(4)]$
Item 7: $l = f \times P(6)$
Item 3: $m = f \times [1 - P(6)]$
Item 6: $n = g \times P(3)$
Item 2: $o = g \times [1 - P(3)]$

Note: *P(t)* is the probability that item *t* is answered correctly given a specific trait level base on the 3PL item response model.

Figure 1. Item Selection at Each Stage of a CAT

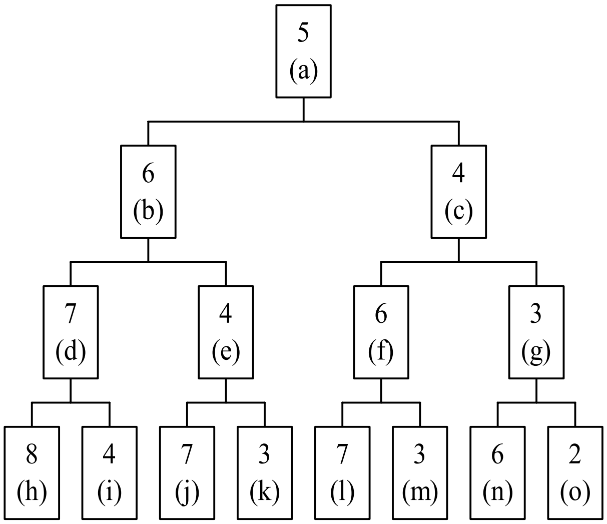

Table 2. Item exposure rates

Item # (i)	Item Exposure Rate (r_i)
1	0.0
2	o
3	$g + k + m$
4	$c + e + i$
5	a
6	$b + f + n$
7	$d + j + l$
8	h
9	0.0
10	0.0
Sum	4.0

Table 3. Descriptive Statistics of the Item Parameters of an ACT-Math Item Pool

Item Parameters	N	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
a	360	0.9688	0.3237	0.2800	2.3700	0.7700	1.4130
b	360	0.3983	1.1221	-3.4300	2.9400	-0.4480	0.3680
c	360	0.1852	0.0865	0.0300	0.5000	1.0630	1.5710

Table 4. Chi-square Statistic

Theta	-3	-2	-1	0	1	2	3
S100	87.91	0.97	0.17	0.40	0.40	0.33	0.06
S500	4.76	0.08	0.07	0.12	0.06	0.05	0.06
S1K	1.19	0.07	0.04	0.04	0.02	0.08	0.02

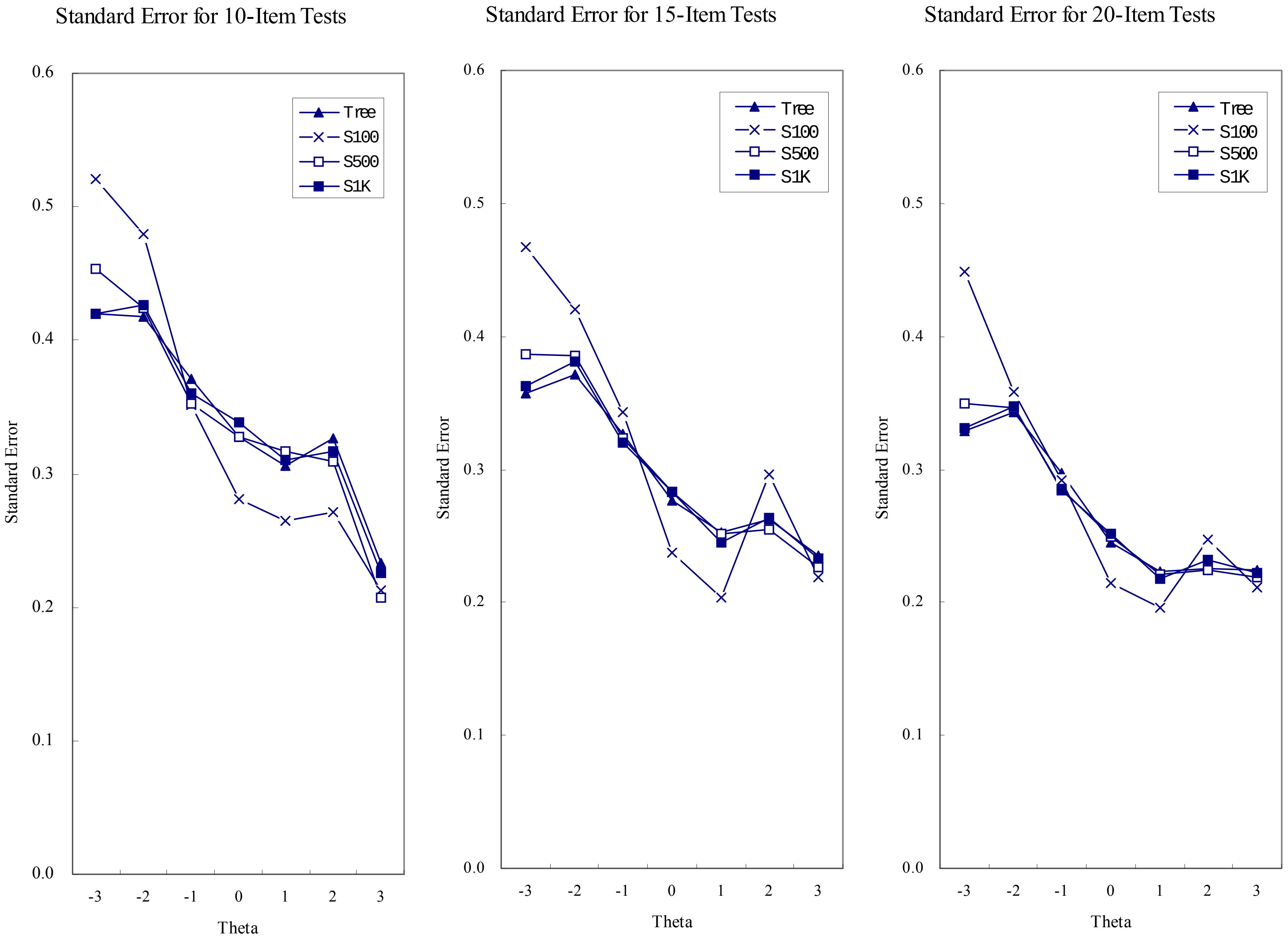

Figure 2. Standard Error

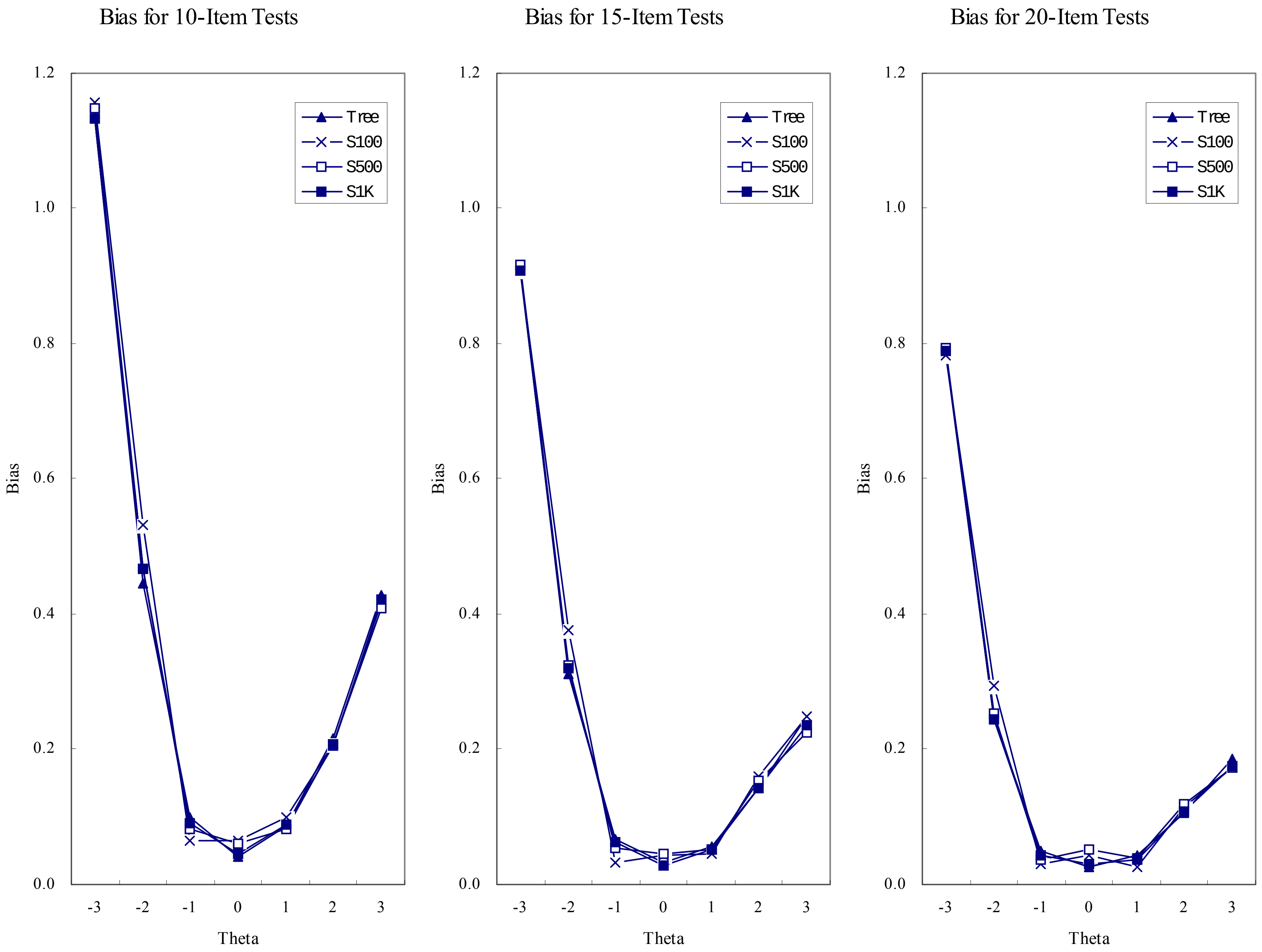

Figure 3. Bias

References

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

Veerkamp, W. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavior Statistics, 22,* 203-226.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Stroud, A. H. & Sechrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.