Journal of Computerized Adaptive Testing

Volume 1 Number 5

December 2013

A Comparison of Four Methods for Obtaining Information Functions for Scores From Computerized Adaptive Tests With Normally Distributed Item Difficulties and Discriminations

Kyoko Ito and Daniel O. Segall

DOI 10.7333/1312-0105088

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 1, Number 5, December 2013
DOI 10.7333/1312-0105088
ISSN 2165-6592

A Comparison of Four Methods for Obtaining Information Functions for Scores From Computerized Adaptive Tests With Normally Distributed Item Difficulties and Discriminations

Kyoko Ito and Daniel O. Segall
Defense Manpower Data Center

A simulation study compared four methods of estimating information for maximum-likelihood estimates (MLE) from fixed-length computerized adaptive tests (CAT): those by (1) Lord (1980), (2) Segall, Moreno, and Hetter (1997), (3) Ito, Pommerich, and Segall (2009a), and (4) Ito and Segall (2010), referred to, respectively, as "local," "quasi-local," "global-slope," and "conditional-averaging-on-θ" (CA-θ). A 900-item bank was constructed using operational three-parameter logistic item parameter estimates [i.e., *a* (discrimination), *b* (difficulty), and *c* (pseudo-guessing)] such that the *a* and *b* parameters were as normally distributed as possible. Test length and the number of simulees at each θ point were varied. Generally, information functions for the quasi-local and global-slope methods were very close to those for the reference local method. However, those for the CA-θ method were distinctly different, which might be explained conceptually. These findings were considerably different than those from previous studies that had used a skewed item bank, and illustrated how bank characteristics, including distributions of item difficulties, discriminations, and ability-bank (mis)matches, could influence which methods would yield similar score information functions.

Computerized adaptive testing (CAT) often utilizes the concept of "information" in multiple ways: during CAT sessions (e.g., when selecting the next item to administer to a single examinee), and outside of actual CAT sessions (e.g., when evaluating the quality of a CAT item bank to be used for numerous examinees). This study dealt with the latter situation. A paper-and-pencil (PP) test based on item response theory (IRT) is typically evaluated in terms of score information and the associated standard error function, which are, respectively, measures of precision and measurement error. Similarly, a CAT bank should be assessed with regard to its score

precision. As in the case of PP tests, such information can be used to ensure that CAT score precision is adequate for an ability range where most examinees are located and that alternate CAT banks are comparable in score precision.

Although only one general method seems commonly employed to obtain a unidimensional score information function for a static set of items (e.g., Equation 1), at least four methods have been suggested and/or researched for unidimensional CAT. They take different approaches to the issue of how to aggregate score information from various CAT tests administered using the items in a single CAT bank. Just as different equating or linking methods could produce varying results when the same set of scores was used, different aggregate score information functions might result from the four methods, even when all the remaining variables are held constant. Additionally, their differences might be systematically affected by some extraneous factors, such as bank features. A large quantity of research findings comparing equating and scale transformation procedures has helped researchers interpret linking or equating results in a more informed manner, and similarly, comparative knowledge of the CAT score information methods should be valuable when a researcher wishes to make intelligent evaluations of CAT information functions reported by other researchers. Prior to the current study, however, little research had explored major characteristics and properties of the four methods while carefully controlling one of the extraneous factors—bank characteristics.

Methods for Computing CAT Information Functions

The methods[1] for computing CAT score information functions through simulation are: (1) the local method by Lord (1980); its two offshoots, (2) the quasi-local method by Segall, Moreno, and Hetter (1997); (3) the global-slope method by Ito, Pommerich, and Segall (2009a, 2009b); and (4) the conditional-averaging-on-θ (CA-θ) method (Ito & Segall, 2010).

The first three methods are based on Birnbaum's (1968) formula for score information. The score information for any test score y (i.e., Fisher information) is defined as

$$I\{\theta, y\} \equiv \frac{\left(\frac{d}{d\theta}\mu_{y|\theta}\right)^2}{Var(y \mid \theta)} = \left(\frac{\frac{d}{d\theta}\mu_{y|\theta}}{SE(y \mid \theta)}\right)^2, \quad (1)$$

where θ is the ability or other trait. When unsquared, the numerator denotes the slope of the regression of score y on θ, while the denominator is the standard error of measurement (*SEM*) of y for a given θ. Note that the regression might be non-linear. Score y can be computed using various scoring methods, including unweighted and weighted summing of item raw scores and IRT scoring.

The local method (Lord, 1980). Equation 1 is primarily intended for scores from traditional non-adaptive tests where all examinees take the same set of items. For scores from CATs that are tailored to examinees (i.e., everyone takes a potentially different set of items), Lord (1980, p.157) provided a formula for computing information functions. It involves simulation and is

[1]The names of the methods were taken from recent literature and are not necessarily the names that were used in their original publications.

based on the conditional mean (m) and variance (s^2) of the final θ estimates for simulees at each of equally spaced θ levels. Specifically,

$$I\left\{\theta,\hat{\theta}\right\} \approx \left(\frac{m(\hat{\theta}\,|\theta_{+1}) - m(\hat{\theta}\,|\theta_{-1})}{(\theta_{+1} - \theta_{-1})} \times \frac{1}{s(\hat{\theta}\,|\,\theta_0)} \right)^2, \tag{2}$$

where $\theta_{-1}, \theta_0,$ and θ_{+1} denote three successive levels of θ, and $\hat{\theta}$ may be obtained with any estimation method. Translated into Birnbaum's (1968) formulation, the first term is the slope, while the denominator of the second term is the *SEM*. Note that this method requires the knowledge of the true θs as well as the $\hat{\theta}$s, and therefore can be implemented only in simulation. This method is referred to here as the *local* method, because the slope is based on two θ points directly adjacent to the θ of interest, i.e., $\theta_{\pm 1}$. As a result, the slope tends to vary from one θ point to another.

The quasi-local method (Segall, Moreno, & Hetter, 1997). The local method tends to produce uneven and unstable information functions. The information functions from Equation 2 can be smoothed by increasing the number of successive θ points to include θ_{-2} and θ_{+2}, as follows:

$$I\left\{\theta,\hat{\theta}\right\} \approx \frac{\left[\frac{m(\hat{\theta}\,|\,\theta_{+1}) + m(\hat{\theta}\,|\,\theta_{+2})}{2} - \frac{m(\hat{\theta}\,|\,\theta_{-1}) + m(\hat{\theta}\,|\,\theta_{-2})}{2} \right]^2}{\left[\frac{\theta_{+1} + \theta_{+2}}{2} - \frac{\theta_{-1} + \theta_{-2}}{2} \right]^2 \left[\frac{1}{5} \sum_{k=-2}^{+2} s(\hat{\theta}\,|\,\theta) \right]^2} \tag{3}$$

$$= \frac{25\left[m(\hat{\theta}\,|\theta_{+2}) + m(\hat{\theta}\,|\theta_{+1}) - m(\hat{\theta}\,|\theta_{-1}) - m(\hat{\theta}\,|\theta_{-2}) \right]^2}{(\theta_{+2} - \theta_{+1} - \theta_{-1} - \theta_{-2})^2 \left[\sum_{k=-2}^{+2} s(\hat{\theta}\,|\,\theta_k) \right]^2}, \tag{4}$$

where $\theta_{-2}, \theta_{-1}, \theta_0, \theta_{+1}, \theta_{+2}$ denote five successive levels of θ. The method for estimating CAT information functions using Equation 4 is referred to as the *quasi-local* method, because it is still based on a small section of the θ scale, but more than three points as in the local method.

The global-slope method (Ito et al., 2009a, 2009b). The global slope method also requires simulation and estimates the slope of Equation 2 with the ordinary least-squares (OLS) regression of $\hat{\theta}$ on θ over the entire θ range. The *SEM* is computed in the same way as in Equation 2, and no smoothing is performed. This method assumes that a single linear line fits the data at hand reasonably well over the θ range; therefore, verifying the appropriateness of applying a single linear line should be part of the method. This third method is referred to as the *global-slope* method, because the slope is global and constant for all θ points, although the *SEM* is estimated for each θ point, as in the local method.

The conditional averaging method (Ito et al., 2010). In a CAT administration there can be as many information functions as the number of unique sets of items, because different examinees tend to take different sets of items from an item bank. The conditional averaging method averages these individual information functions conditionally on either θ or $\hat{\theta}$. That is, for a

group of simulees/examinees that share the same θ (or $\hat{\theta}$) value, a score information function is computed for each of them based on the set of items administered, and the information value that corresponds to the θ (or $\hat{\theta}$) value is found. Such information values are then averaged to obtain an unweighted average for the given θ (or $\hat{\theta}$) value, and the process is repeated for all reasonably finite values of θ (or $\hat{\theta}$). The conditional averaging method is a departure from the typical way to estimate the amount of score information as a function of the slope of $\hat{\theta}$ on θ and the *SEM* of $\hat{\theta}$ for a given θ (e.g., Birnbaum, 1968), but it is a way to obtain an aggregate information function for alternate CAT forms that utilize the items in the same item bank.

Simply stated, the first three methods (i.e., local, quasi-local, and global-slope) are based on Birnbaum's fundamental formula for Fisher score information—that is, squared slope divided by variance (i.e., squared *SEM*); the methods involve only the final θ estimates and do not involve score information functions for the individual CAT tests. The fourth method takes an entirely different approach by computing score information functions for all the examinees and then averaging score information values at the examinee's true (or estimated) θ values. In terms of computational or programming complexity, they are very comparable.

On one hand, an advantage of the local method is its faithfulness to Birnbaum's original formula. On the other hand, its major disadvantage, as the name suggests, is its narrow focus on two adjacent θ points. As a result, any "noise" or outliers that almost always exist locally in the data can substantially influence each slope-*SEM* pair and lead to uneven score information functions. Another disadvantage, albeit minor, is that no score information is calculable for θ points on either end, because the slope is based on three θ points.

The quasi-local method improves on the local method in the sense that it smoothes out the jaggedness by widening the local method's narrow focus. As such, the resulting quasi-local score information functions show general trends very well. However, some people might consider them to be approximations or trend lines.

A strength of the global-slope method is that it removes the impact of local noise and outliers in the data from the slope, which is derived from the entire θ range. Thus, a researcher who prefers a linear relationship between true and estimated θs to be based on total data, as opposed to a few θ points, might favor this method. Although the resulting score information functions still tend to be uneven, the jaggedness can largely be eliminated by increasing the number of simulees (which also applies to the local method). To the extent the linearity is imperfect, however, either the local or quasi-local method should serve the researcher better.

It should be noted that information functions obtained by Lord's local method would be considerably higher than a simple average of all the individual Fisher information functions from CAT sessions. Straight averaging would result in a substantially lower aggregate information function because each individual information function would be peaked at or around the examinee's $\hat{\theta}$, which varies across the θ range.

The Current Study in Relation to Previous Research

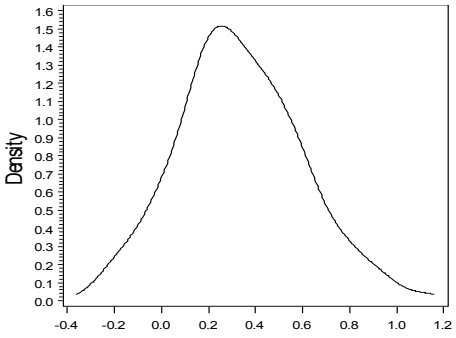
Considering the pros and cons of the four methods, a crucial question is how similar or dissimilar their score information functions are and under which conditions. At least the three studies mentioned above made such comparisons. The first two studies (Ito, Pommerich, & Segall, 2009a , 2009b) compared three methods—the local, quasi-local, and global-slope methods. The

third study (Ito & Segall, 2010) added a fourth method (CA-θ). All three studies were identical in terms of the item bank, fixed-length CAT implementation rules for item selection and exposure control, the use of a uniform simulee θ distribution, and the variables that were manipulated. They differed only in one aspect: type of score examined. The first and third studies employed maximum-likelihood estimation (MLE), while the second study used Bayesian modal estimation (BME) with a N(0,1*)* prior. Ito et al. (2009a) found the differences in MLE information functions among the three methods, on average, to be very small if the number of simulees at each θ point was at least 500 and test length was 15 items or greater.

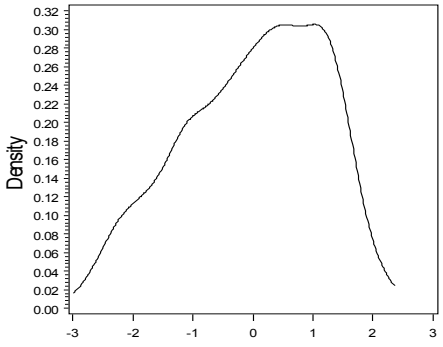
The results from the first study, however, were not replicated in the second study (Ito et al., 2009b). In the second study, which used BME, score information functions for the local and quasi-local methods were similar, while those for the global-slope method were distinctly different. The source of the differences was the shape of the item bank used. As shown in Figures 1a and 1b, the item bank, which was a random subset of existing operational CAT item banks and therefore realistic, was skewed with too few very difficult items and too many moderately difficult items. The lack of very difficult items in the bank caused θ estimates on the high end to shrink toward the prior mean on shorter CAT tests. Interestingly, only the global-slope method produced aggregate score information functions that signaled the reduced precision for high-ability simulees.

Figure 1. Distribution of the Ln(Discrimination) and Difficulty Parameters Used in the Ito et al. Studies (2009a, 2009b, 2010)

a. Ln(Discrimination)



b. Difficulty



Using the item parameters for the items administered in the simulated CAT sessions in the first study, the third study (Ito et al., 2010) explored how MLE information functions from the CA-θ method were related to those from the local method. The study observed that at the largest N_k (2,000) and test length (60 items), their information functions were extremely similar to those from the local method. It was tentatively concluded that information functions from the CA-θ method and the local method might be asymptotically[2] identical. However, the possibility existed that this result was an artifact of the simulation conditions, including the characteristics of the item bank.

[2] "Asymptotically" here means both the number of simulees at each θ point and test length.

Thus, the prior studies highlighted the importance of controlling simulation conditions other than those frequently manipulated, such as test length and both sample and bank size. One such infrequently explored condition is item bank characteristics. The reasons why the current study focused on them were twofold. First, the first and second studies revealed that item bank characteristics could lead to markedly different score information functions. Second, the asymptotic equivalence of the local and CA-θ methods observed in the third study needed to be shown to be replicable under different conditions. Ideally, a simulation study should vary a condition in a systematic way, preferably based on knowledge about typical values of the condition. However, there seems to be little information on typical characteristics of operational CAT item banks. Lacking this information, the present study strove to establish a baseline by creating an item bank whose item difficulties and discriminations were approximately or largely normally distributed. Although such an item bank might not always be achievable in practice, this bank shape might be deemed to be desirable in the area of ability-achievement assessments considering that abilities are typically relatively normally distributed. Note that the only conditional difference between the present and past studies was the item bank—all the other aspects remained identical. To the extent that results from the present study differ from those in the earlier studies, the generalizability of the earlier results would be called into question, while the impact of item bank structure on score information would be underscored.

Method

Source of the Item Parameters

The item parameters for this study's simulation originated in existing banks for fixed-length multiple-choice CAT tests that have been operationally administered to a large population for more than twenty years. The IRT parameters for all items in the banks had been estimated using the three-parameter logistic (3PL) model,

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]}, \qquad (i = 1 \ldots n) \tag{5}$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, a_i, b_i, and c_i are the discrimination, difficulty, and pseudo-guessing parameters of item i, and D is a scaling constant, 1.7.

From the operational CAT item banks, items were randomly selected for replacement with the goal of creating a 900-item CAT bank in which the b parameters were as normally distributed as possible. The bank size was selected to accommodate the maximum test length and exposure rate used in the study (discussed further below). Table 1 presents descriptive statistics for the 3PL item parameters as well as the natural logarithm of the a parameter [i.e., Ln(a)]. The distributions of the item parameters for the 900 items are plotted in Figures 2a–2d. Compared with the b parameters in the 600-item bank that was used in the preceding three studies (e.g., Ito et al., 2009a), the b parameters in the current 900-item bank were distributed symmetrically around 0. Similarly, the distribution of the Ln(a) parameters was more bell-shaped and centered around 0 than it was in the earlier studies. Perhaps more importantly, the current b parameter distribution extended beyond ±3.0 unlike the b parameters in the previous 600-item bank whose maximum

value was 2.35 when the true θs ranged between ±3.0. The normal probability plots (available in the Supplementary Data file) suggested that the present *b* parameters substantially followed the standard normal distribution and that the Ln(*a*) parameters were largely normally distributed. With a mean of 0.21 and a standard deviation of 0.08, the distribution of the *c* parameters does not appear atypical.

Table 1. Descriptive Statistics of the *a*, Ln(*a*), *b*, and *c* Parameters for the Item Bank

Parameter	*N*	Mean	*SD*	Min.	Max.
a	900	1.099	0.382	0.412	3.761
Ln(*a*)	900	0.041	0.321	-0.886	1.325
b	900	0.011	1.003	-3.663	3.586
c	900	0.206	0.084	0.017	0.500

Response Generation

The simulated CAT for this study matched the actual operational implementation of the CAT testing program as much as possible, including the use of Sympson-Hetter (1985) exposure control, the 3PL model, and Lord's (1980) maximum information item selection method. The operational CAT procedure computes provisional $\hat{\theta}$s using Owen's (1969) Bayesian method, and final $\hat{\theta}$s are computed using the Bayesian modal method with a N(0, 1) prior. Also, the initial $\hat{\theta}$ is set to 0.0 for all examinees. However, this operational CAT scoring procedure was modified for this study, which focused on MLE. The simulation to produce exposure control parameters also employed MLE.

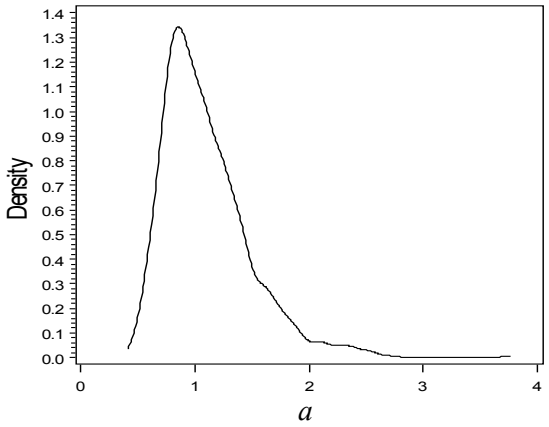
Using the item parameters described above, item responses were generated for a fixed number of simulees at each of 31 equally-spaced θ points between ±3.0. See the Simulation Conditions section below for the number of simulees. Specifically, for a given θ, the first item was selected to have the highest information value. The response to the first item (or any subsequent item) was assigned a 1 if $P_i(\theta)$ was greater than a single-precision pseudo-random number from a uniform distribution between 0 and 1, and 0 otherwise. The initial $\hat{\theta}_{\mathrm{MLE}}$ was set to ±3.0 depending on the response to the first item. $\hat{\theta}_{\mathrm{MLE}}$ remained ±3.0 until a sequence of "all incorrect" or "all correct" was broken, after which θ was freely estimated until the final θ estimate, which was then bounded to ±3.0. The second and all subsequent items were selected such that each had the maximum information value based on the latest $\hat{\theta}_{\mathrm{MLE}}$, had not yet been administered to the simulee, and satisfied the exposure control threshold.

Throughout the simulation, the item parameters were treated as "true" item parameters that were known, eliminating the possibility of differences due to item parameter estimation (e.g., different calibration programs and estimation algorithms). That is, all information functions used in the current study were computed using the true item parameters summarized in Table 1 and Figure 2.

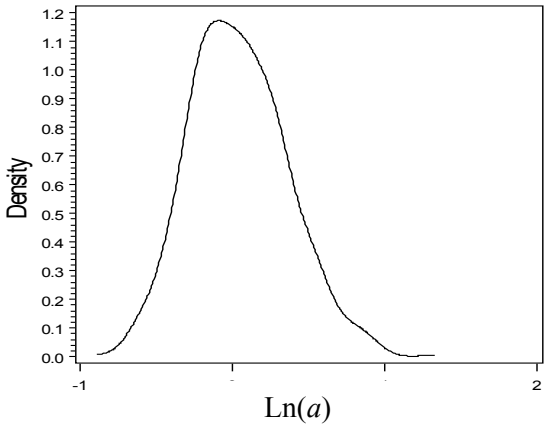
In generating item exposure control parameters using the Sympson-Hetter procedure, separate simulations were conducted for each of the different simulated test lengths (see Simulation Conditions below), where the maximum exposure rate was set to equal the exposure rate observed in typical PP administrations of the operational tests, i.e., 0.167.

Figure 2. Distribution of Item Parameter Estimates in the 900-Item Bank and its Fisher Information Function

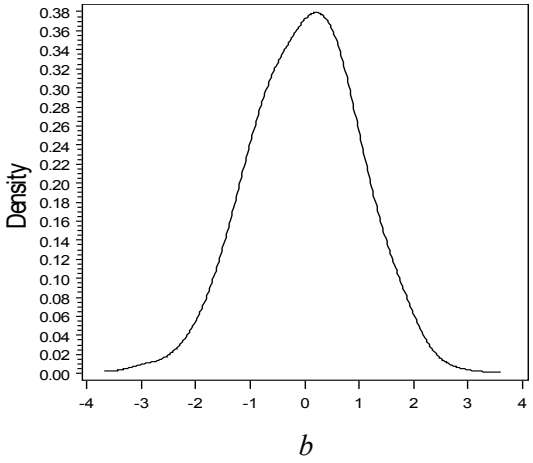
a. *a* Parameter



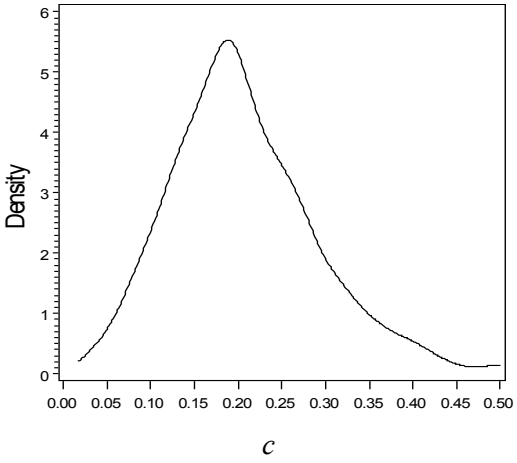
b. Ln(*a*)



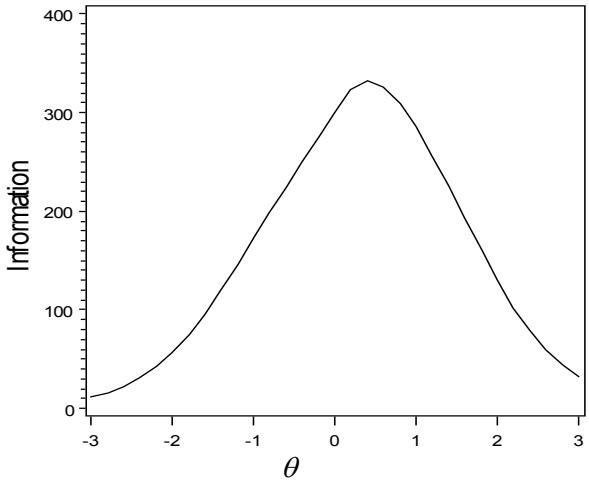
c. *b* Parameter



d. *c* Parameter



e. Fisher Information for the Bank



Simulation Conditions

The manipulated simulation conditions were as follows:

1. Test length: 15, 20, 30, and 60 items. A majority of the operational tests of interest are longer than 15 items in length. The test length of 60 items was intended to simulate an asymptotic (i.e., long test) condition.
2. Number of simulees at each equally-spaced θ point (N_k): 100, 500, 1,000, and 2,000. Total sample size (N) was determined by the number of simulees at each equally-spaced θ point. For example, under the condition where N_k = 100, sets of responses were generated at each of the 31 θ points, so total N was 3,100 simulees.

The study employed a fully-crossed design (4 × 4), so the number of simulees at each θ point (N_k) varied at each of the four test lengths. Each combination of test length and N_k was replicated 10 times.

Results

Appropriateness of the Global-Slope Method

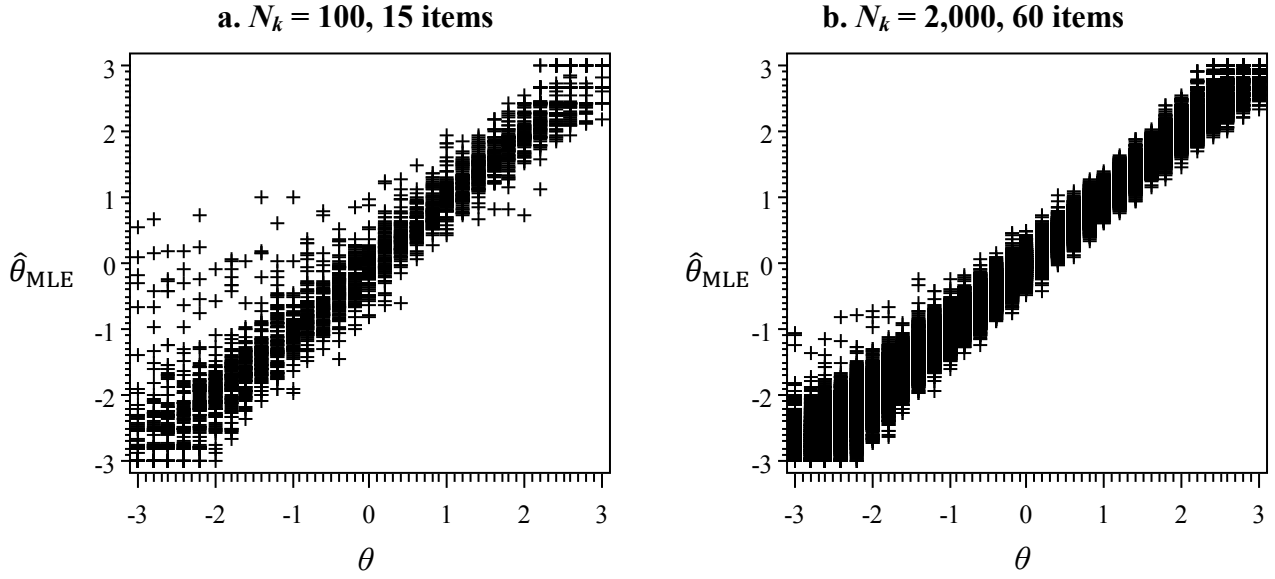
Use of the global-slope method assumes that it is appropriate to fit a linear line throughout the entire range of θ and $\hat{\theta}_{\text{MLE}}$. Figure 3 presents sample scatterplots of θ and the final $\hat{\theta}_{\text{MLE}}$ from Replication 1, respectively, for the shortest test (15 items) with the smallest N_k (100), and the longest test (60 items) with the largest N_k (2,000). The remaining replications produced substantially similar scatterplots, which are available in the Supplementary Data file. The figures seem to strongly indicate the appropriateness of using the OLS estimator to obtain a global linear slope as opposed to a non-linear line (e.g., splines). For additional information, Table 2 shows mean R^2s (averaged without Fisher z transformation) for the regression of final $\hat{\theta}_{\text{MLE}}$ on θ. Regardless of N_k, the R^2s were .95 for the 15-item test, .97 for the 20-item test, .98 for the 30-item test, and .99 for the 60-item tests. These R^2s seem to confirm the appropriateness of using the OLS estimator in all conditions.

Comparisons of Information Functions Between the Local and Each of the Remaining Three Methods

The comparisons were first made graphically. The MLE information functions from all four methods from Replication 1 are plotted in Figure 4. The MLE information functions from the other nine replications were very comparable and are available in the Supplementary Data file. A few noticeable features include:

1. The information functions from the local and global-slope methods tend to be jagged, particularly at N_k = 100, while those from the quasi-local and CA-θ methods are smooth regardless of N_k or the number of items. The cragginess of the local and global-slope information functions should be attributed to their common element, i.e., the *SEM* that is computed for each θ point and left unsmoothed. Note that the unevenness largely disappeared at N_k = 2,000 (Figures 4d, 4h, 4l, and 4p).

Figure 3. Sample Scatterplots of Final $\hat{\theta}_{\text{MLE}}$ and θ (Replication 1)



2. As indicated by the relatively small differences between the local and either the quasi-local or global-slope method, the information functions for these three methods are substantially similar, and the smooth quasi-local information functions run through the uneven local and global-slope information functions.
3. The CA-θ information functions tend to hover discernibly above the information functions for the other three methods. This is consistent with the numerical results, i.e., the root-mean-squared-differences (*RMSD*s) and relative root-mean-squared-differences (%*RMSD*s). Those for the CA-θ method were almost always the largest among the three methods (Tables 3 and 4).

Table 2. R^2 for the Ordinary Least-Squares Regression of Final $\hat{\theta}_{MLE}$ on θ, Averaged Over 10 Replications

	Test Length			
N_k	15	20	30	60
100	0.950	0.966	0.980	0.989
500	0.951	0.967	0.981	0.989
1,000	0.951	0.967	0.981	0.989
2,000	0.952	0.967	0.980	0.989

The score information functions in Figure 4 were then evaluated in terms of *RMSD*s in information between the local method and each of the other three methods, averaged over the 31 θ points and ten replications. The *RMSD* was defined as

$$RMSD = \sqrt{\frac{\sum_{k=1}^{L} d_k^{\,2}}{L}}, \tag{6}$$

where d_k is the difference in information between the two methods at a given θ level (the information value for the local method was subtracted from that of each other method), and L is the number of θ levels. The score information values yielded by the local method are given in the Appendix, and descriptive statistics for the *RMSD*s are given in Table 3.

The *RMSD* was computed to provide an index of how large, on average, the differences in absolute magnitude were. They generally were small, varying between 0.83 for the quasi-local method with the largest N_k of 2,000 and the shortest test of 15 items, and 7.30 for the CA-θ method with the smallest N_k of 100 and the longest test of 60 items. Notable observations include:

1. The *RMSD*s tended to be slightly larger for the global-slope method than for the quasi-local method when test length was relatively short (i.e., 15 or 20 items), but slightly smaller with longer test lengths (i.e., 30 and 60 items).
2. The CA-θ method unfailingly produced the largest *RMSD*s among the three methods.
3. For the quasi-local and global-slope methods, the *RMSD* decreased as N_k increased but increased as the test became longer. The *RMSD* for the CA-θ method did not seem to have clear trends.
4. The general pattern that the standard deviations (*SD*s) of the *RMSD*s were considerably small relative to the means seemed to justify the number of replications used (10).

Because it was difficult to assess whether the *RMSD*s in Table 3 were large or small, relative differences (i.e., %*RMSD*s) were obtained by dividing each *RMSD* by the maximum amount of information observed for the local method under a given simulation condition. The maximum amounts of information for the local method are reported in Table 4, while geometric means of the %*RMSD*s over ten replications are presented in Table 5. The maximum information values for the local method ranged from about 18 to 23 for the 15-item test, 25 to 34 for the 20-item test, 36 to 42 for the 30-item test, and 57 to 66 for the 60-item test.

The %*RMSD*s for the quasi-local and global-slope methods in Table 5 varied between 3.1% (for the global-slope method with the largest N_k of 2,000 and the longest length of 60 items) and 11.9% (for the quasi-local method with the smallest N_k of 100 and the shortest length of 15 items). Those for the CA-θ method, ranging from 5.5% to 20.4%, were often remarkably larger. As with the *RMSD*s, the %*RMSD*s tended to be marginally larger for the global-slope method than for the quasi-local method when test length was relatively short (i.e., 15 or 20 items), but slightly smaller with longer test lengths (i.e., 30 and 60 items). Unlike *RMSD*, however, the %*RMSD* fairly steadily declined as a function of either N_k or test length for the quasi-local or global-slope methods. Once again, %*RMSD*s for the CA-θ method did not seem to have recurring patterns.

Figure 4. Information Functions for MLEs From Simulated Tests of 15 to 60 Items and N_k = 100, 500, 1,000, and 2,000 for Replication 1

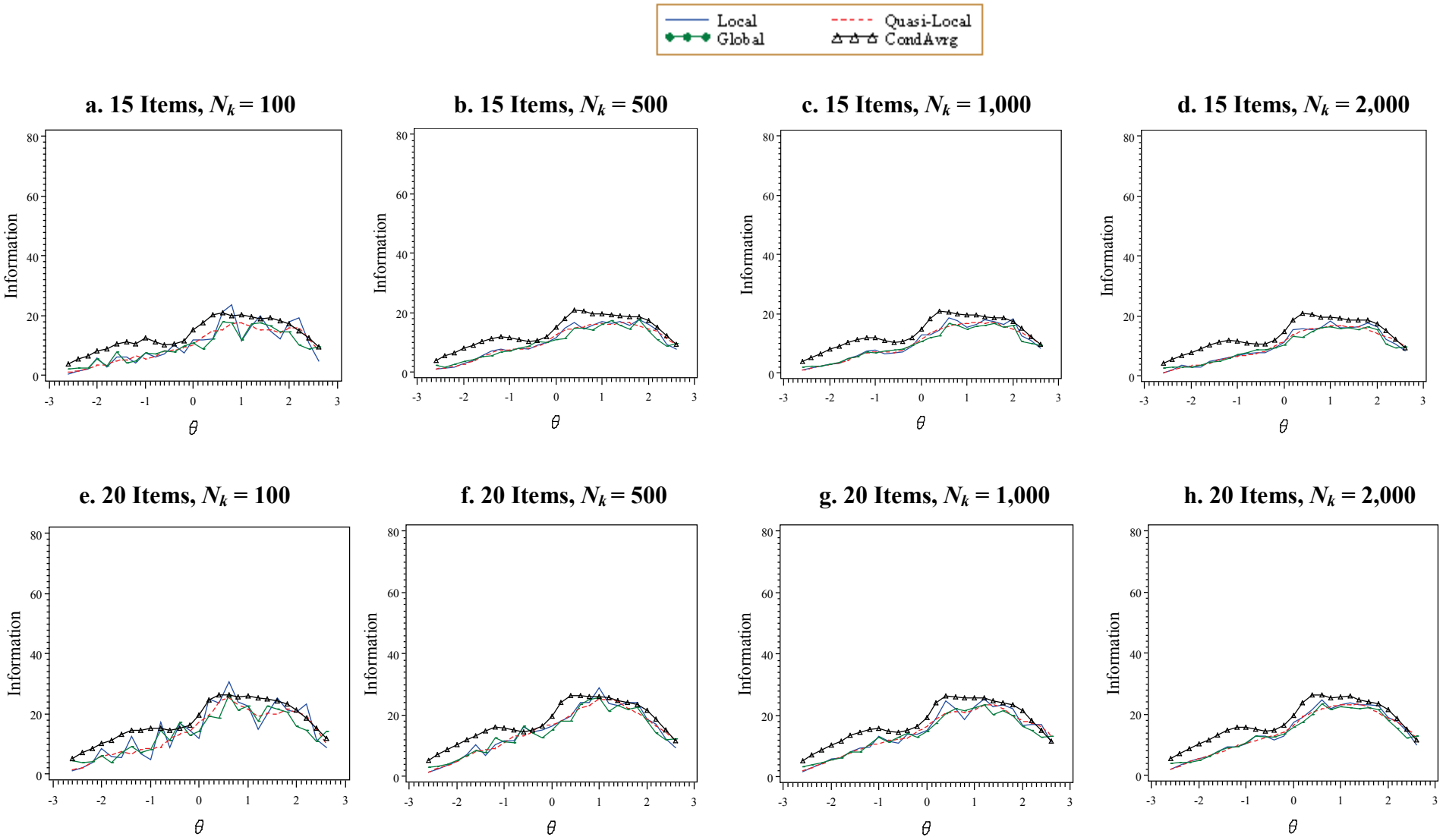


Figure 4 (continued). Information Functions for MLEs From Simulated Tests of 15 to 60 Items and N_k = 100, 500, 1,000, and 2,000 for Replication 1

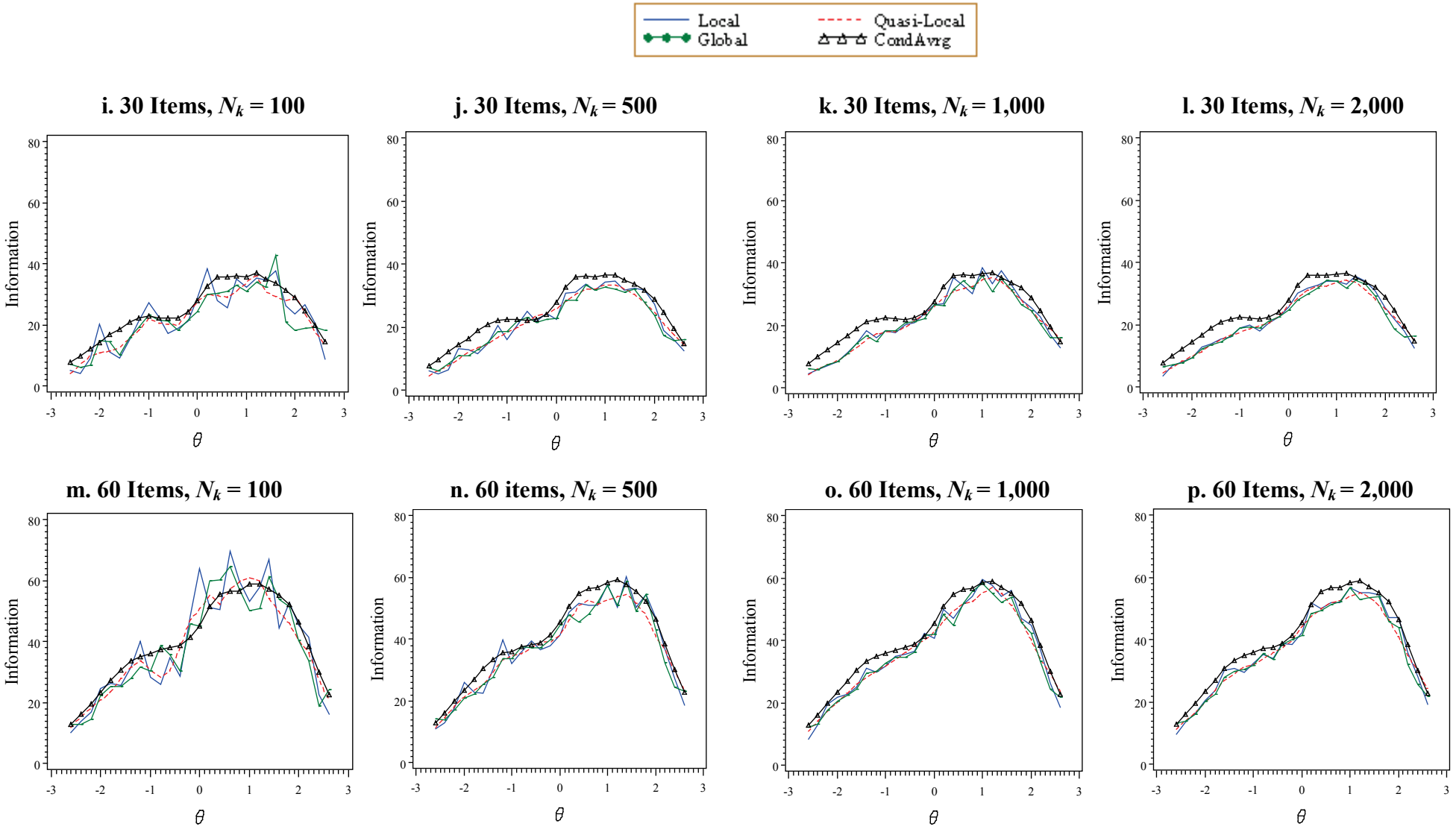


Table 3. *RMSD* of Information for Tests of 15, 20, 30, and 60 Items for the Quasi-Local, Global-Slope, and CA-θ Methods

N_k and Statistic	Quasi-Local				Global-Slope				CA-θ			
	15	20	30	60	15	20	30	60	15	20	30	60
N_k = 100												
Mean	2.97	2.64	4.27	6.51	2.64	3.02	3.56	4.83	4.59	5.11	5.29	7.30
SD	0.38	0.40	0.70	1.69	0.40	0.41	0.48	1.12	0.44	0.52	0.66	1.41
Min	3.67	3.27	5.44	8.98	3.27	3.66	4.18	6.64	5.20	6.26	6.49	9.66
Max	2.51	2.17	3.10	4.01	2.17	2.36	2.79	2.78	3.96	4.56	4.49	5.42
N_k = 500												
Mean	1.25	1.47	2.17	3.08	1.47	1.68	1.96	2.44	3.86	3.83	3.92	3.77
SD	0.36	0.12	0.37	0.35	0.12	0.17	0.23	0.31	0.23	0.19	0.27	0.28
Min	2.17	1.61	2.88	3.64	1.61	2.03	2.28	2.92	4.31	4.26	4.24	4.09
Max	0.95	1.27	1.64	2.56	1.27	1.50	1.57	1.99	3.52	3.57	3.45	3.18
N_k = 1,000												
Mean	0.99	1.24	1.73	2.37	1.24	1.50	1.66	1.91	3.78	3.74	3.46	3.25
SD	0.16	0.08	0.12	0.23	0.08	0.17	0.12	0.24	0.11	0.12	0.16	0.25
Min	1.25	1.39	1.91	2.66	1.39	1.77	1.78	2.35	3.98	3.90	3.72	3.52
Max	0.74	1.14	1.48	2.03	1.14	1.22	1.40	1.54	3.60	3.51	3.23	2.78
N_k = 2,000												
Mean	0.83	1.10	1.39	1.97	1.10	1.27	1.51	1.75	3.68	3.66	3.40	3.16
SD	0.09	0.05	0.10	0.19	0.05	0.06	0.06	0.15	0.06	0.07	0.11	0.20
Min	1.00	1.17	1.51	2.28	1.17	1.38	1.67	1.98	3.81	3.77	3.53	3.40
Max	0.71	1.03	1.16	1.72	1.03	1.18	1.46	1.56	3.58	3.57	3.14	2.78

Table 4. Test Length for the Maximum Local Information Method

	Test Length			
N_k	15	20	30	60
100	22.79	34.47	42.14	65.97
500	18.79	28.05	36.26	60.98
1,000	18.49	25.33	37.51	60.08
2,000	18.45	24.62	36.28	57.06

Table 5. Percent *RMSD* of the Maximum Local Information Method

	Test Length			
Method and N_k	15	20	30	60
Quasi-Local Method				
$N_k = 100$	11.88	10.38	9.98	9.24
$N_k = 500$	6.34	5.64	5.70	5.07
$N_k = 1{,}000$	5.25	5.05	4.63	4.01
$N_k = 2{,}000$	4.55	3.80	3.86	3.45
Global-Slope Method				
$N_k = 100$	10.53	8.86	8.36	6.88
$N_k = 500$	7.62	6.31	5.17	4.01
$N_k = 1{,}000$	6.68	5.90	4.46	3.22
$N_k = 2{,}000$	6.12	5.18	4.22	3.06
CA-θ Method				
$N_k = 100$	18.43	15.06	12.42	10.51
$N_k = 500$	20.05	14.44	10.40	6.22
$N_k = 1{,}000$	20.32	14.74	9.29	5.50
$N_k = 2{,}000$	20.42	14.93	9.47	5.55

Discussion and Suggestions for Future Research

The quasi-local information functions were very similar to the reference local-method information functions. This was not in the least unexpected, for the only difference between the two methods is that the quasi-local method applies smoothing to information functions from the local method. The global-slope information functions were also substantially similar to the reference information functions. This is also logical, because the relationship between θ and the final $\hat{\theta}_{\text{MLE}}$ appeared generally linear even though the final $\hat{\theta}_{\text{MLE}}$s were bounded by ±3.0, and as a result, the distinction between "local" versus "global" slope was immaterial. Relatively speaking, on shorter tests the information functions by the quasi-local method tended to be the most similar to the

reference functions, while on longer tests those by the global-slope method tended to be the most comparable.

In contrast to the results for the quasi-local and global-slope methods, the CA-θ method produced information functions that were generally but consistently higher and differed the most from those by the local method. The plots of the MLE information functions showed that the CA-θ MLE information functions tended to be somewhat higher than those for the other three methods nearly throughout the θ range. Additionally, the differences in information between the local and CA-θ methods had a tendency to exhibit different patterns as a function of test length or N_k than did those between the local and either the quasi-local or global-slope methods. It is not clear whether the averaging conditionally on θ of individual score information functions from a CAT session is bound to yield aggregate information functions that are different from those by the local, quasi-local, and global-slope methods when a normally distributed item bank is used. Conceptually, as noted above, it seems to make sense to think of the latter three methods as one class (i.e., "slope-divided-by-*SEM*" methods) from which the CA-θ method is excluded.

Above all, this study has highlighted that bank characteristics affect score information functions. The MLE results from the present study using a bank with virtually normally-distributed *a* and *b* parameters were markedly dissimilar to those from the previous studies that utilized a bank with asymmetrically distributed *b* parameters. For example, the Ito et al. study (2009a) with a skewed distribution of *b* parameters showed the quasi-local MLE information functions to be unmistakably lower in the center than the local functions; such a recurring feature entirely disappeared in the present study with more normally distributed *b* parameters. As another instance, the MLE information functions by the CA-θ method in the Ito et al. study (2010) seemed asymptotically equivalent to those by the local method, but this finding was not replicated in the current study that used a different bank. In short, it is not only score precision that can fluctuate considerably and sometimes unexpectedly depending on what kinds of items are available in a bank for CAT tests, but also which methods might produce similar and dissimilar score information functions.

A substantial amount of CAT research on item banks has so far focused on item exposure, security, and the efficiency of bank usage (e.g., van der Linden & Veldkamp, 2004; Cheng, Chang, Douglas, & Guo, 2009). It seems that the current study, as well as its predecessor studies, have demonstrated that evaluating item banks from the perspective of score information might also provide illuminating findings about strengths and deficiencies of CAT banks. Additionally, a subsequent extension of the current research to the BME has confirmed the finding reported by Ito et al. (2009b) that an ability-bank mismatch has a more salient impact on BME than on MLE. Therefore, it seems beneficial to include an ability-bank match-mismatch as an aspect of an item bank as did some past research (e.g., Gorin, Dodd, Fitzpatrick, & Shieh, 2005).

Suggestions for Future Research

Recommendations based on the current line of research include the following.

1. Verify that the relationship between θ estimates and true θ is nearly perfectly linear. The lack of linearity might signal something anomalous.
2. If linearity is verified, any of the local, quasi-local, and global-slope methods can be used to obtain CAT score information functions. The advantages and disadvantages are outlined in the Methods for Computing CAT Information Functions section.

3. Use an N_k of at least 2,000 to generate smoother functions, regardless of the method used.
4. If there is reason to believe that the audience (e.g., lay people) would be confused or distracted by slight irregularities in the information functions, use the quasi-local method to obtain smoothed functions.
5. At least until further data are available, the CA-θ method should not be considered a method for generating aggregate CAT score information functions.
6. Check to determine whether examinees and items are reasonably well-matched. If they are, the choice of MLE or BME does not seem to matter; otherwise, MLE seems more robust to the lack of alignment than does BME.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Cheng, Y., Chang, H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints—balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, 69, 35–49. CrossRef

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. CrossRef

Ito, K., Pommerich, M., & Segall, D. O. (2009a). An evaluation of a new procedure for computing information functions for scores from computerized adaptive tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Ito, K., Pommerich, M., & Segall, D. O. (2009b). An evaluation of a new procedure for computing Bayesian scores from computerized adaptive tests. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. McLean, VA: Graduate Management Admission Council. Available from http://iacat.org/biblio

Ito, K., & Segall, D. O. (2010). An evaluation of a new procedure for obtaining information functions for maximum-likelihood scores from computerized adaptive tests: Conditional averaging on theta. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (RB-69-92). Princeton, NJ: Educational Testing Service.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 117–130). Washington, DC: American Psychological Association. CrossRef

Sympson, J.B., & Hetter, R.D. (1985). Controlling item exposure rates in computerized adaptive

tests. Paper presented at the Annual Conference of the Military Testing Association. San Diego, CA.

van der Linden, W.J., & Veldkamp, B.P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*, 273–291. CrossRef

Supplementary Data

The Supplementary Data file for this article contains the following data:

- Normal probability plots for the *b* and Ln(a) parameters
- Scatterplots of final $\hat{\theta}_{\text{MLE}}$ and θ for each of 10 replications
- Information functions for each of 10 replications

This file can be requested from the Editor, djweiss@umn.edu.

Acknowledgments

Practically all of the work for this article was completed while the first author was an employee of the Human Resources Research Organization (HumRRO). This article is her thanks for the wonderful five years at HumRRO. She is also especially grateful for the valuable guidance and input from the JCAT editors, particularly Gage Kingsbury.

Author Address

Kyoko Ito, Defense Manpower Data Center, 400 Gigling Rd., Seaside, CA 93955-6771. U.S.A.
Email kyoko.ito.civ@mail.mil.

Appendix

Table A-1. Score Information Values for the Local Method, for N_k = 100, 500, 1,000, and 2,000, and 15- and 20-Item Tests

	15 items				20 items			
θ	100	500	1,000	2,000	100	500	1,000	2,000
-2.6	1.20	0.98	1.03	1.08	2.55	2.03	2.04	1.96
-2.4	2.09	1.64	1.70	1.84	3.42	2.96	2.89	3.14
-2.2	2.13	2.68	2.73	2.77	4.96	4.53	4.12	4.17
-2.0	4.27	3.20	3.34	3.41	4.56	5.40	5.61	5.23
-1.8	3.76	4.08	3.71	3.78	6.14	6.69	6.51	6.52
-1.6	4.37	4.77	4.54	4.68	10.19	7.87	7.62	7.54
-1.4	7.72	5.61	5.79	5.58	8.39	8.62	8.99	8.95
-1.2	6.15	6.92	6.63	6.48	9.52	10.68	10.52	10.15
-1.0	7.68	7.08	7.21	6.98	10.67	11.73	11.18	10.47
-0.8	6.89	6.97	7.41	7.60	13.79	11.08	11.20	11.70
-0.6	7.01	8.48	7.44	7.76	12.17	12.16	11.61	12.01
-0.4	10.07	9.14	8.46	8.56	13.77	13.23	13.05	12.54
-0.2	9.72	9.32	9.95	9.50	15.49	14.70	13.78	13.95
0.0	10.60	11.71	11.57	11.61	15.39	16.88	16.85	16.80
0.2	15.09	13.80	13.12	14.03	19.81	19.35	20.69	19.59
0.4	16.58	14.47	15.19	15.51	23.88	22.14	22.28	22.31
0.6	15.84	16.49	16.05	16.44	24.14	23.26	22.88	23.27
0.8	19.70	16.80	16.42	16.04	24.52	23.07	23.22	22.80
1.0	17.78	17.30	17.15	17.00	23.06	22.72	22.57	22.78
1.2	15.69	16.60	16.76	16.46	26.95	24.07	23.56	23.46
1.4	17.45	17.52	17.07	17.21	22.55	23.76	23.61	23.72
1.6	17.70	16.63	17.34	17.47	23.81	22.51	22.86	22.98
1.8	17.66	17.13	17.09	17.15	22.48	22.11	21.83	22.07
2.0	17.90	16.01	16.07	15.96	19.85	19.92	19.10	19.56
2.2	15.06	13.59	13.86	13.07	17.87	17.04	16.23	16.92
2.4	10.05	10.95	10.82	10.24	12.43	13.04	14.61	13.80
2.6	7.99	8.32	7.89	8.38	10.55	10.26	10.93	10.44

Table A-2. Score Information Values for the Local Method, for N_k = 100, 500, 1,000, and 2,000, and 30- and 60-Item Tests

	30 items				60 items			
θ	100	500	1,000	2,000	100	500	1,000	2,000
-2.6	5.47	4.94	4.31	4.08	9.95	10.50	10.01	10.28
-2.4	6.42	6.78	6.61	6.41	13.58	14.04	13.41	14.13
-2.2	7.09	8.15	8.55	8.49	19.36	17.50	19.00	18.08
-2.0	12.21	10.46	9.90	9.67	21.24	21.47	21.95	20.73
-1.8	13.08	11.84	11.67	11.99	27.06	24.71	23.95	24.23
-1.6	13.98	13.36	13.85	13.71	28.37	27.98	27.89	28.24
-1.4	19.35	15.69	15.76	15.71	27.67	31.39	31.03	30.85
-1.2	16.85	17.77	17.50	17.84	31.42	32.63	31.50	32.28
-1.0	21.60	17.05	18.43	18.62	35.65	34.53	34.11	33.12
-0.8	18.62	19.33	19.12	18.45	34.77	35.24	34.34	34.64
-0.6	21.10	21.25	19.28	18.73	35.19	36.65	35.87	35.87
-0.4	22.98	20.45	21.37	21.27	39.57	37.00	37.49	38.29
-0.2	23.18	22.62	22.40	22.43	41.14	39.81	39.79	39.08
0.0	25.07	25.20	25.78	24.41	45.48	43.39	43.74	43.06
0.2	30.49	30.18	30.03	29.89	48.22	46.95	48.05	48.45
0.4	31.95	32.94	33.95	32.30	50.05	52.45	49.86	50.73
0.6	34.14	31.36	33.05	32.65	54.50	52.33	52.70	52.69
0.8	32.29	32.02	32.28	33.90	55.43	54.38	53.79	52.94
1.0	32.20	34.32	34.58	33.82	55.61	56.29	55.70	55.28
1.2	36.68	35.57	34.64	34.56	56.52	56.94	56.75	55.40
1.4	36.05	33.69	33.55	34.29	57.56	57.02	55.26	54.66
1.6	32.44	31.44	34.08	32.70	51.44	53.67	53.65	53.87
1.8	32.09	30.40	29.81	29.62	48.41	49.18	49.07	48.56
2.0	25.16	26.02	26.56	27.46	47.56	44.89	45.28	43.49
2.2	23.61	21.07	22.58	22.44	38.15	35.70	35.85	35.99
2.4	19.05	17.62	17.48	17.43	28.42	26.71	27.38	27.38
2.6	12.41	13.67	13.12	13.30	20.79	19.21	19.89	19.78