Graduate
Management
Admission
Council®

An Evaluation of a New Procedure for Computing Information Functions for Bayesian Scores From Computerized Adaptive Tests

Kyoko Ito
Human Resources Research Organization
Mary Pommerich
Defense Manpower Data Center
Daniel O. Segall
Defense Manpower Data Center

Presented at the Item and Pool Development Paper Session, June 3, 2009

Abstract

Using Bayes modal θ estimates with a N(0,1) prior, a simulation study evaluated how a new procedure for estimating information for scores from CAT administrations (i.e., the global-slope method) compared to two existing procedures (i.e., the local and quasi-local methods). Test length and the number of simulees at each θ point were varied. A fundamental premise of the global-slope approach is a single linear relationship between θ and $\hat{\theta}$ over the entire θ range. The linearity assumption was found to hold for tests with 30 and 60 items, while tests with 10 and 15 items yielded nearly linear, but slightly S-shaped scatterplots of θ and $\hat{\theta}_{\text{BME-N}(0,1)}$. However, due to a confounding factor caused by the pool used, the study was unable to assess the impact of the minor violation of the assumption on the information functions of the shorter tests. On the tests of 30 and 60 items, however, the three methods produced relatively similar information functions, particularly when there were at least 500 simulees at each θ point.

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

Ito, K., Pommerich, M., & Segall, D. (2009). An evaluation of a new procedure for computing information functions for Bayesian scores from computerized adaptive tests. 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

Kyoko Ito, HumRRO, 40 Ragsdale Dr., Suite 150 , Monterey, CA 93940.
Email: kito@humrro.org

An Evaluation of a New Procedure for Computing Information Functions for Bayesian Scores From Computerized Adaptive Tests

Computerized adaptive testing (CAT) requires a pool of items that can yield reliable scores for a range of examinees without compromising security. One way to evaluate CAT item pools is to compare them in terms of their information functions. For example, of two equal-size pools, the one with more discriminating items (and therefore higher information) for an ability range where examinees are located is more desirable. Also, if one intends to construct alternate item pools, they should be comparable in their information functions.

As described by Birnbaum (1968), the score information for any test score *y* is defined as:

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

where θ is ability. The numerator denotes the slope of the regression of score *y* on θ, while the denominator is the standard error of measurement 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 number-correct scores and item response theory (IRT) scoring. One of the most frequently used IRT (θ) estimators is the maximum likelihood estimator (MLE). Lord (1980) states that when the weights for item scores are optimal as a function of θ, score *y* becomes a MLE ($\hat{\theta}_{\mathrm{MLE}}$), and that the asymptotic (i.e., for a sufficiently long test) information function for score $\hat{\theta}_{\mathrm{MLE}}$ on a test of *n* items has a special name, i.e., the test information function. The MLE information function is expressed as:

$$I(\hat{\theta}_{MLE}) = \frac{1}{Var(\hat{\theta}_{MLE}\,|\,\theta)} = \sum_{j=1}^{n} \frac{P_j'^2}{P_j Q_j}, \tag{2}$$

where P_j is the logistic function for item *j*, $Q_j = 1 - P_j$, and P'_j is the first derivative of P_j. Theorem 5.3.2 in Lord (1980) states that $I(\hat{\theta}_{\mathrm{MLE}})$ "is an upper bound to the information that can be obtained by any method of scoring the test" (p. 71). Interestingly, the information function for the Bayes modal estimator with a normal prior (BME $_{normal}$), $I(\hat{\theta}_{BME-normal})$, includes a positive additive term as follows:

$$I(\hat{\theta}_{\text{BME-normal}}) = I(\hat{\theta}_{\mathrm{MLE}}) + \frac{1}{\sigma^2}, \tag{3}$$

where σ^2 is the variance of the posterior distribution.

The above formulas for the MLE or BME $_{\mathrm{normal}}$ score information, however, are primarily used for scores from traditional non-adaptive tests where all examinees take the same set of items. For scores from CAT tests that are tailored to examinees (i.e., everyone takes a potentially different set of items), Lord (1980, p.157) provides a formula for computing information functions. It involves simulation and is based on the conditional mean and variance of the final scores for simulees at each of equally-spaced θ levels. Specifically,

$$I\{\theta,\hat{\theta}\} \approx \frac{\left[m(\hat{\theta}\,|\,\theta_{+1}) - m(\hat{\theta}\,|\,\theta_{-1})\right]^2}{(\theta_{+1}-\theta_{-1})^2\, s^2(\hat{\theta}\,|\,\theta_0)}, \tag{4}$$

where *m* is the conditional mean, s^2 is the variance of the final scores, and θ_{-1}, θ_0, and θ_{+1} denote three successive levels of θ. 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 procedure will be referred to as the *local method*, because the slope is based on two θ points directly adjacent to the θ of interest.

However, the local method tends to produce uneven and unstable information functions. The information functions can be smoothed by increasing the number of successive θ points to include θ_{-2} and θ_{+2} as follows:

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

$$= \frac{25\left[m(\hat{\theta}|\theta_{+2}) + m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1}) - m(\hat{\theta}|\theta_{-2})\right]^2}{(\theta_{+2}-\theta_{+1}-\theta_{-1}-\theta_{-2})^2 \left[\sum_{k=-2}^{+2} s(\hat{\theta}|\theta_k)\right]^2}, \tag{6}$$

where $\theta_{-2}, \theta_{-1}, \theta_0, \theta_{+1}, \theta_{+2}$ denote five successive levels of θ (Segall, Moreno, & Hetter, 1997). The procedure for estimating CAT information functions using Equation 5 will be 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.

Yet another method seems possible via simulation if one has sufficient evidence that a single linear line fits the data reasonably well over the θ range— i.e., a method in which the slope of Equation 1 is estimated with ordinary least-squares (OLS) regression across the entire θ range and divided by $s^2(\hat{\theta}|\theta_0)$. This third method will be referred to as the *global-slope method*, because the slope is global and constant for all θ points, although the standard error is estimated for each θ point, as in the local method.

This study compared the new global-slope procedure with the two existing procedures for computing amounts of information from simulated CAT administrations. The type of score employed was the BME with a N(0,1*)* prior ($\text{BME}_{\text{N}(0,1)}$). In comparing the procedures, the study considered such aspects as test length and sample size. Prior to applying the global-slope method, it was evaluated whether it would be appropriate to fit a linear line to the entire range of θ and $\hat{\theta}_{BME-N(0,1)}$.

Method

Source of the Item Parameters

The source of the item parameters for the simulation was nine existing CAT pools of multiple-choice Arithmetic Reasoning (AR) items. These nine pools were developed for use in large-scale operational administrations of a 15-item fixed-length test. The IRT parameters for all items in the nine pools had been estimated using the three-parameter logistic (3PL) model and placed onto a single operational scale. The nine item pools were assembled in such a way as to achieve similar measurement precision and content coverage across the pools. See Sands, Waters, & McBride (1997), DMDC (2008), and DMDC (2009) for more details of the assembly of the AR item pools.

From the nine CAT item pools, items were randomly selected to create a 600-item CAT pool for the simulation. The pool size was selected to accommodate the test length and maximum exposure rate used in the study (discussed further below). Table 1 presents descriptive statistics for the 3PL item parameter estimates (i.e., *a*/discrimination, *b*/difficulty, and *c*/pseudo-chance) as well as the natural logarithm of the *a* [i.e., Ln(*a*)]. The distributions of the item parameter estimates for the 600 items are plotted in Figures 1. Although Table 1 shows that the *b*s were virtually centered around 0, Figure 2 demonstrates a lack of very difficult items in the pool, considering that the true θs ranged between ±3.0 as described in the next section. This, as it turned out, had a noticeable impact on the results for the global-slope method.

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

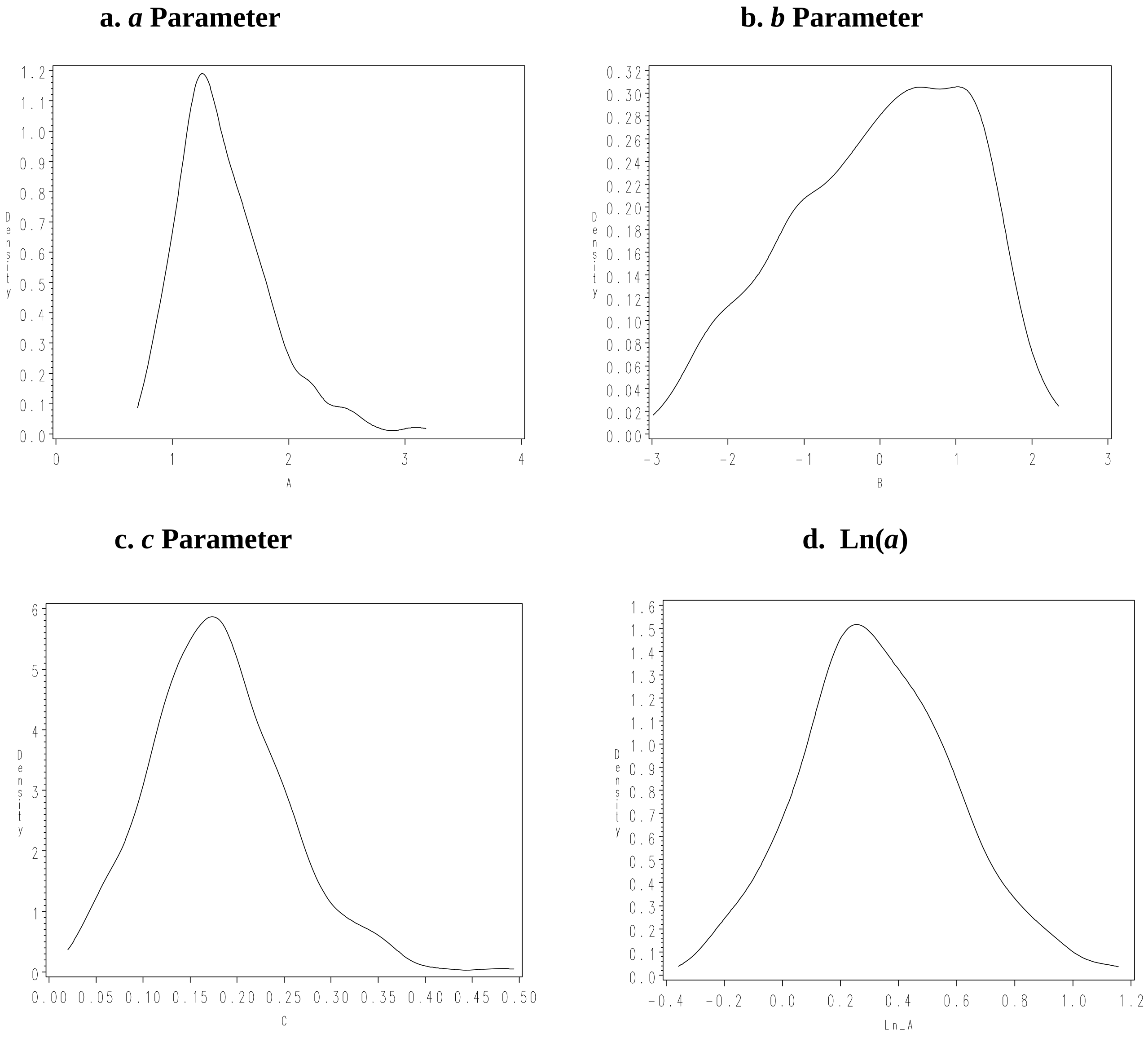
Parameter	*N*	Mean	SD	Min.	Max.
a	600	1.445	0.403	0.700	3.178
Ln(*a*)	600	0.332	0.265	-0.357	1.156
b	600	-0.002	1.173	-2.983	2.350
c	600	0.181	0.072	0.020	0.495

Response Generation

Item responses were generated for a fixed number of simulees at each of 31 equally-spaced θ points between −3.0 and +3.0, employing the 3PL model and the item parameter estimates. The number of simulees at each θ point varied across conditions. Throughout the simulation, these item parameter estimates were treated as “true” item parameter values that were known, eliminating the possibility of differences due to item parameter estimation (e.g., different calibration programs and estimation algorithms). That is, all $\text{BME}_{N(0,1)}$ information functions were computed using the true item parameter values summarized in Table 1 and Figure 1.

The simulated CAT for the study matched the actual operational implementation of the CAT testing program as much as possible, including the use of Sympson-Hetter (1985) exposure control and Lord’s (1980) maximum information item selection procedure. The study employed the scoring procedure used in the operational CAT administration: provisional $\hat{\theta}$s were obtained using Owen’s (1969) Bayesian sequential method and final $\hat{\theta}$s using the Bayes modal method with no limit imposed on the maximum or minimum values. In generating item exposure control parameters via the Sympson-Hetter procedure, separate simulations were conducted for

Figure 1. Distribution of Item Parameter Estimates in the 600-Item Pool



a. *a* Parameter **b. *b* Parameter**

c. *c* Parameter **d. Ln(*a*)**

each of the differing simulated test lengths (see “Simulation Conditions” below), where the maximum exposure rate was set to equal the exposure rate observed in paper-and-pencil administrations of the AR test, i.e., 1/6.

Simulation Conditions

Two factors were manipulated in the simulation: test length and sample size. The manipulated simulation conditions were as follows:

1. Test length: 10, 15, 30, and 60 items.

   The test length of 15 items matched that used in operational CAT administrations of the AR test. 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, N was 3,100 simulees under the $N_k = 100$ condition where 100 sets of responses were generated at each of the 31 θ points.

Each combination of test length and N_k was replicated five times.

Results

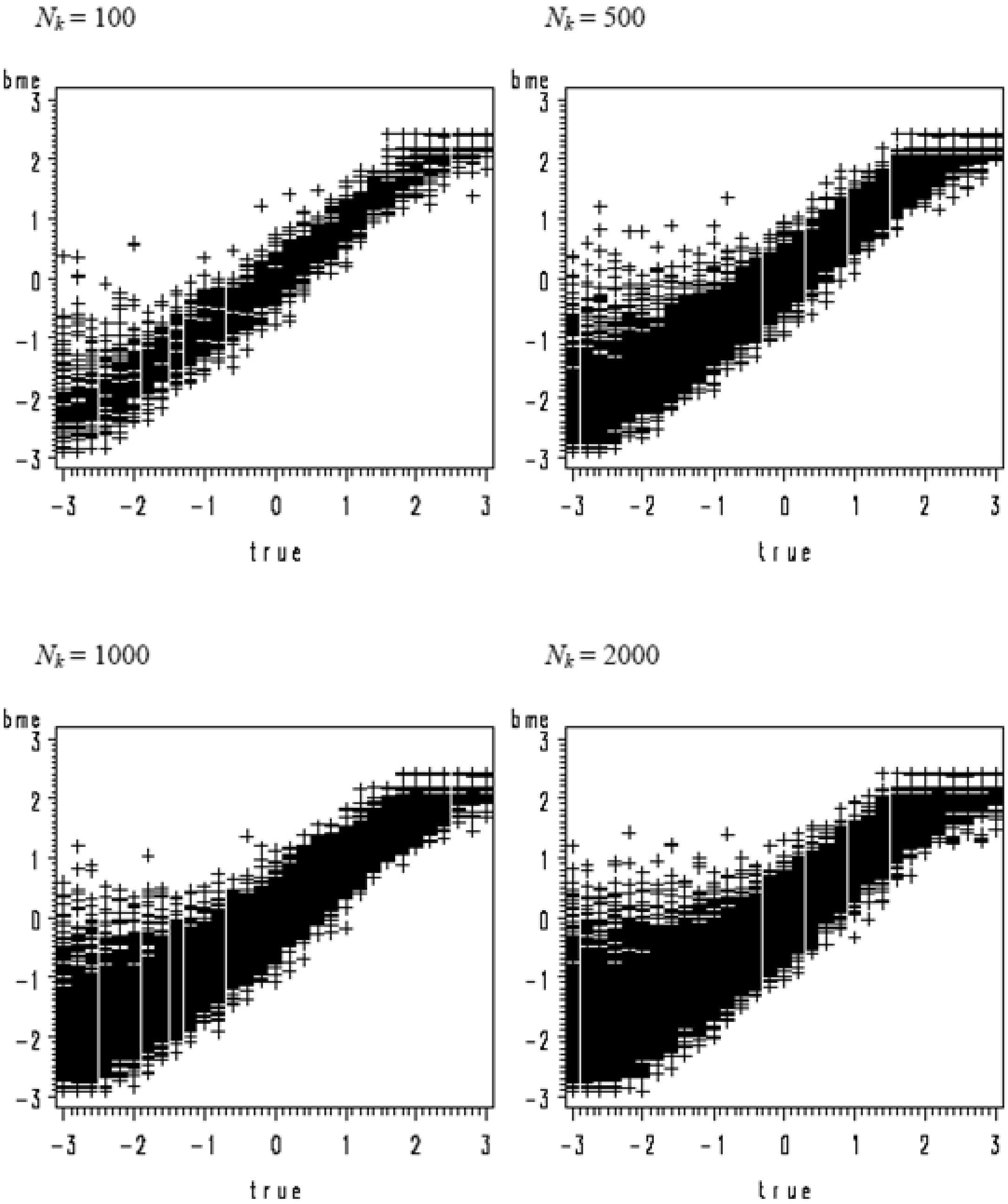
Appropriateness of the Global-Slope Method

Use of the global-slope method assumes that it is appropriate to fit a linear line to the entire range of θ and $\hat{\theta}_{\text{BME-N(0,1)}}$. Figures 2 –5 present scatterplots of θ and final $\hat{\theta}_{\text{BME-N(0,1)}}$ from Replication 1. The remaining replications produced substantially similar scatterplots and are not included here. Figures 2 and 3 for the shorter-length tests seem to suggest that the relationship between θ and $\hat{\theta}_{\text{BME-N(0,1)}}$ is almost linear, but slightly S-shaped. What is troubling about these plots, however, is a gap that is present on the upper-right corner of each plot, indicating that $\hat{\theta}_{\text{BME-N(0,1)}}$ clearly shrank toward the mean of the prior and did not reach 3.0 (the maximum θ value used). This "failing-to-reach-3" feature can be attributed to the lack of more difficult items in the pool that was noted earlier in the Method section.

In contrast, Figures 4 and 5 for the longer tests demonstrate substantial linearity between θ and $\hat{\theta}_{\text{BME-N(0,1)}}$. Additionally, the plots seem to exhibit no shrinkage toward the prior mean in the sense that $\hat{\theta}_{\text{BME-N(0,1)}}$ is obviously reaching 3.0. Furthermore, the R^2s for the regression of final $\hat{\theta}_{\text{BME-N(0,1)}}$ on θ were extremely consistent across all five replications. Regardless of N_k, the R^2 was .94 or .95 for the 10-item test, .96 or .97 for the 15-item test, .98 or .99 for the 30-item test, and .99 for all the 60-item tests.

In sum, the linearity assumption of the global-slope method seemed to be satisfied when test length was longer (e.g., at least 30 items), while it was slightly violated on shorter tests (e.g., 15 or fewer items). More important, the lack of difficult items in the pool caused Bayesian θ estimates on the 10- and 15-item tests to shrink toward the mean of the prior and fail to attain the maximum θ value of 3.0. As can be seen in the next section, this has an implication for the global-slope method. No such shrinkage was observed for the 30- and 60-item tests.

Figure 2. Scatterplots of Final $\hat{\theta}_{\text{BME-N(0,1)}}$ and θ From 10-Item Simulated Tests (Replication 1)



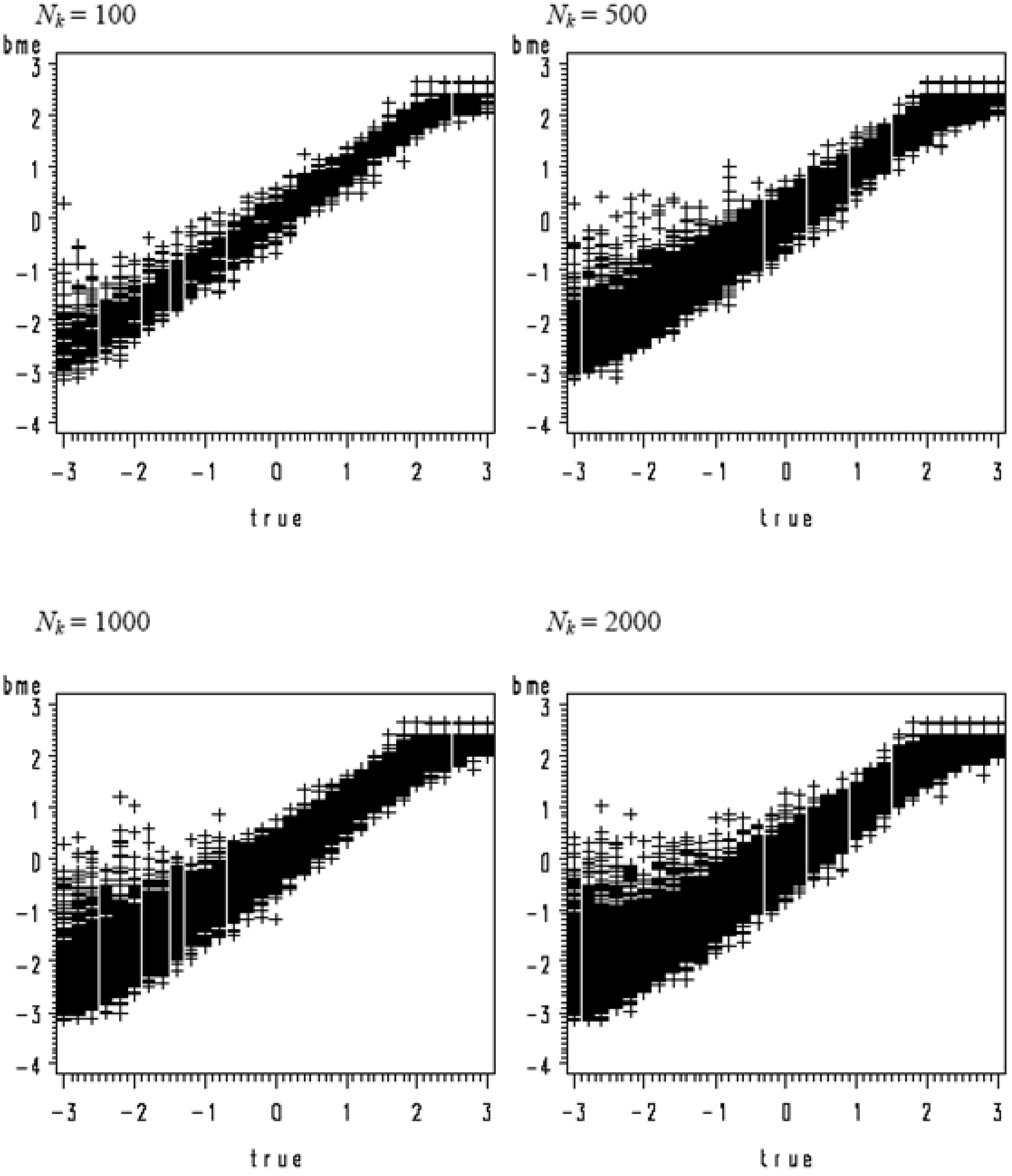
$N_k = 100$

$N_k = 500$

$N_k = 1000$

$N_k = 2000$

Figure 3. Scatterplots of Final $\hat{\theta}_{\text{BME-N}(0,1)}$ and θ
From 15-Item Simulated Tests (Replication 1)



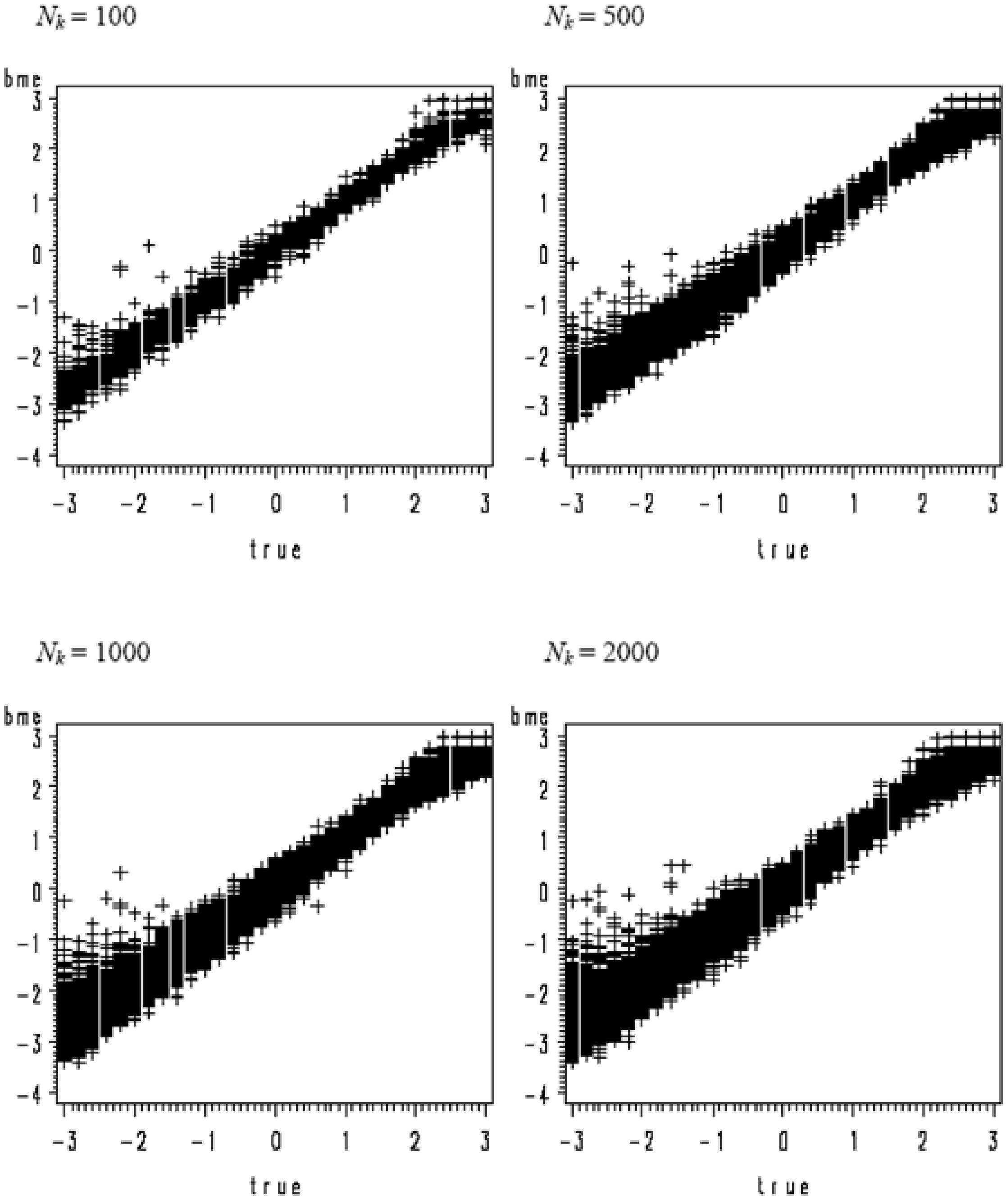
$N_k = 100$

$N_k = 500$

$N_k = 1000$

$N_k = 2000$

Figure 4. Scatterplots of Final $\hat{\theta}_{\text{BME-N}(0,1)}$ and θ From 30-Item Simulated Tests (Replication 1)



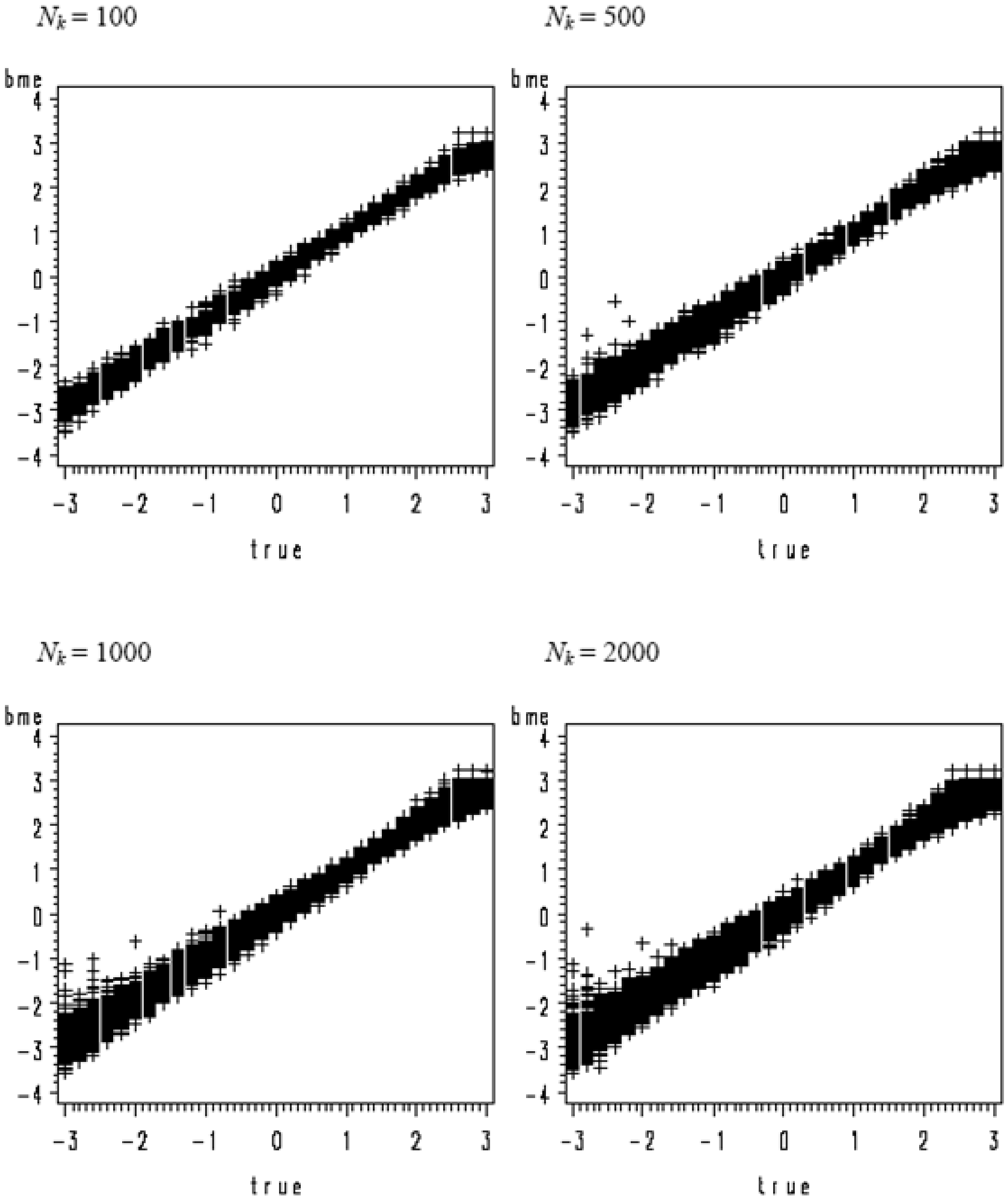
$N_k = 100$

$N_k = 500$

$N_k = 1000$

$N_k = 2000$

Figure 5. Scatterplots of Final $\hat{\theta}_{\text{BME-N(0,1)}}$ and θ
From 60-Item Simulated Tests (Replication 1)



Comparisons of BME Score Information Functions From the Three Approaches

The information functions from the three methods from Replication 1 are plotted in Figures 6 – 9. Once again, these information functions were very similar across all five replications.

Figure 6. Information Functions for BME$_{N(0,1)}$

Scores From 10-Item Simulated Tests (Replication 1)

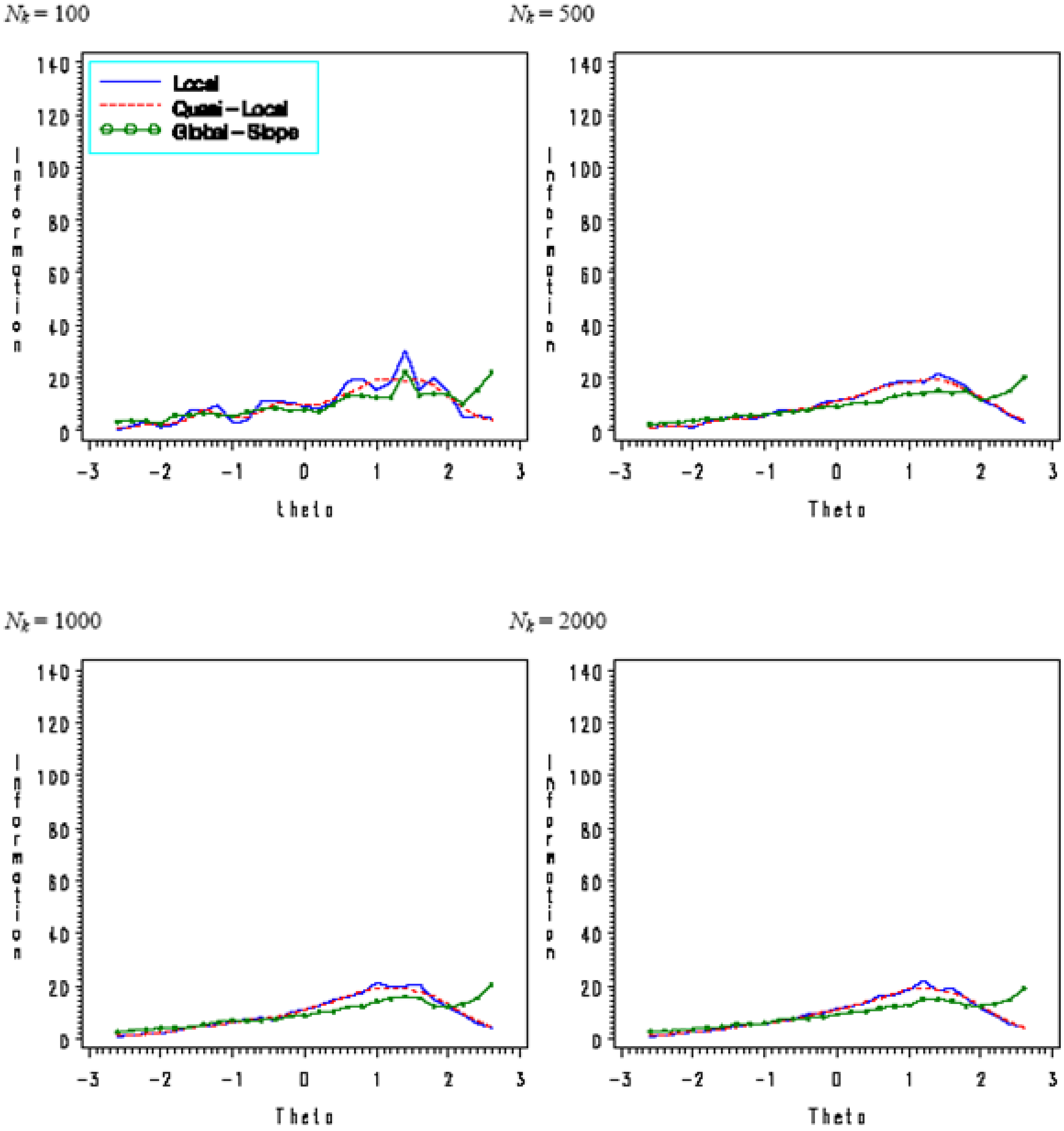
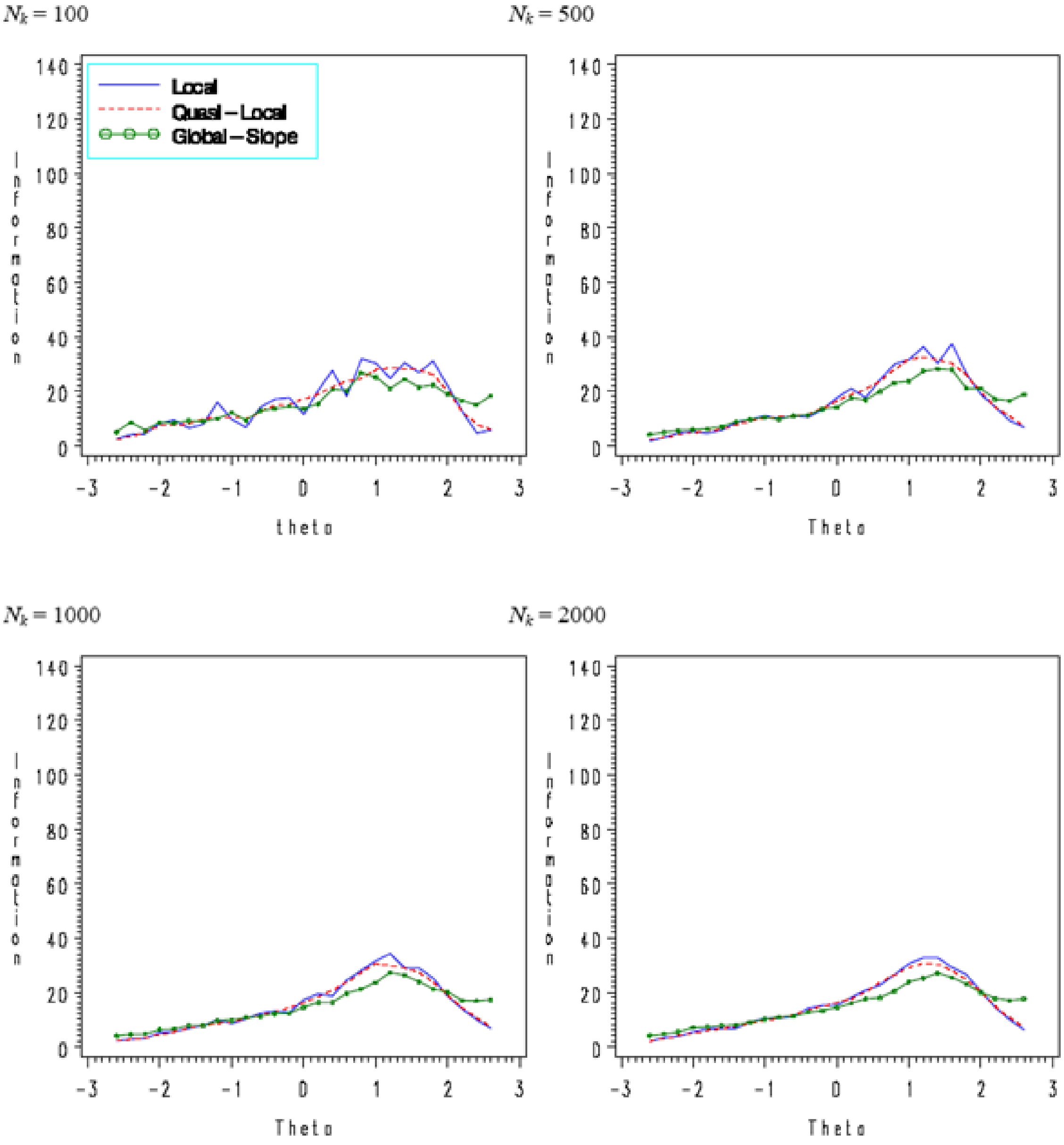


Figure 7. Information Functions for BME$_{N(0,1)}$ Scores From 15-Item Simulated Tests (Replication 1)


$N_k = 100$

$N_k = 500$

$N_k = 1000$

$N_k = 2000$

Figure 8. Information Functions for $\text{BME}_{N(0,1)}$
Scores From 30-Item Simulated Tests (Replication 1)

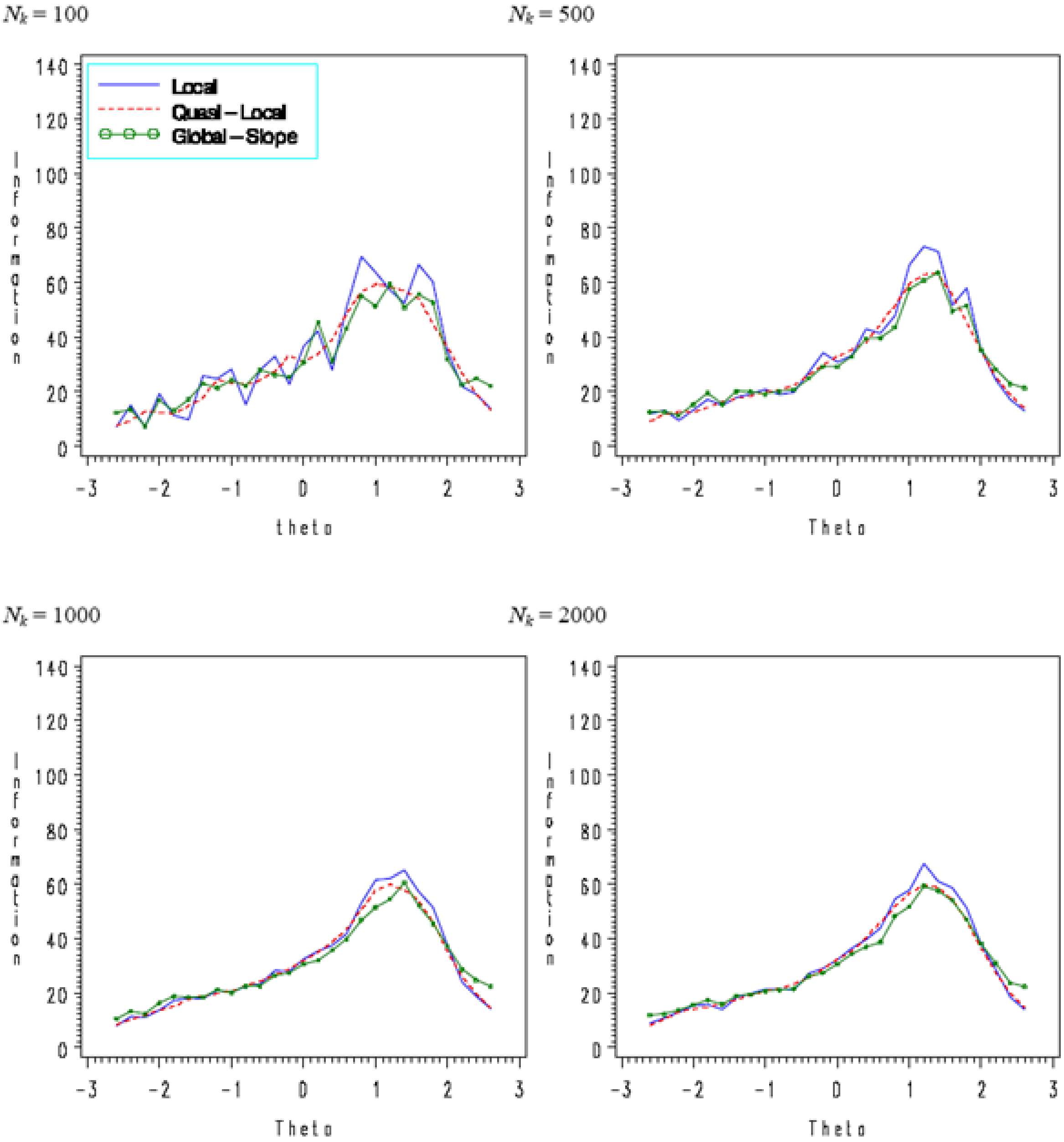
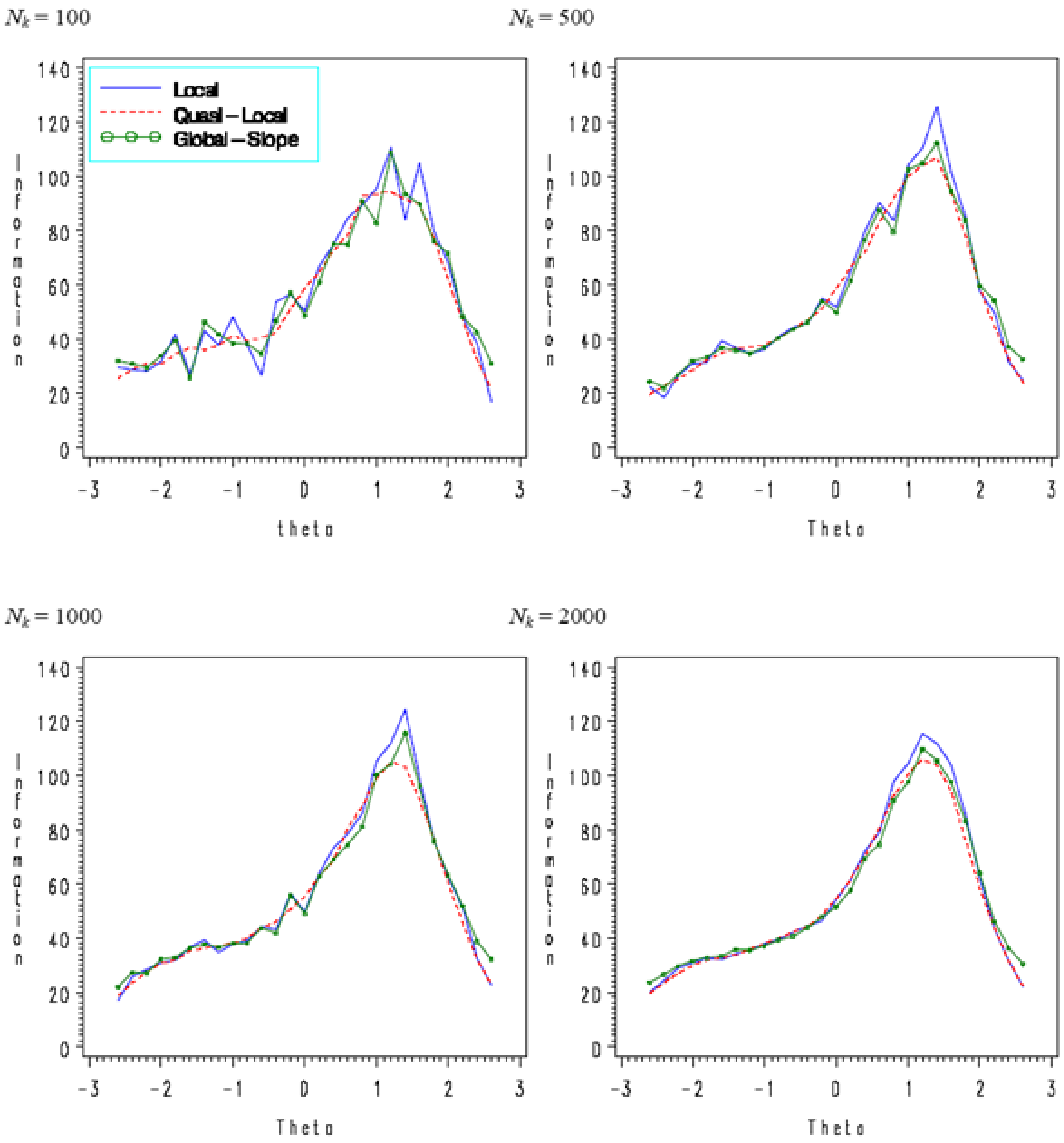


Figure 9. Information Functions for BME$_{N(0,1)}$

Scores From 60-Item Simulated Tests (Replication 1)



On the shorter tests (Figures 6 and 7), the local and quasi-local information functions are very similar, particularly when N_k is 500 or larger. Although the global-slope information functions are fairly comparable on the lower half to those based on the local and quasi-local methods, they clearly differ on the upper half with a distinctive rise on the high end. This rise on the upper end is a result of the "failing-to-reach-3" phenomenon noted earlier (i.e., shrinkage toward the prior mean), which, in turn, is a result of the lack of more difficult items in the pool.

Thus, the insufficient number of difficult items, which is unrelated to the global-slope method, confounded the results. If the pool had contained as many difficult items as easier items, the results for the global-slope method, even on the shorter tests, would plausibly have been more comparable to those from the other existing methods. Since the plots seem sufficiently informative at this point, numerical comparative results will be reported in a future report when the same comparisons are made using a fuller pool. See the Discussion section for a more detailed explanation of why the shortage of difficult items in the pool affected only the global-slope method's results.

In contrast, all three methods produced relatively similar information functions on the 30- and 60-item tests, particularly for N_k of 500 or larger (Figures 8 and 9). Given that the assumption of the global-slope method was met on the longer tests, this came as no surprise. Information functions of the global-slope method seem to have a tendency to be more similar to those of the quasi-local method than to those of the local method.

Discussion

Summary and Conclusions

This study evaluated how a new procedure for estimating information for scores from CAT administrations (i.e., the global-slope method) compared to two existing procedures (i.e., the local and quasi-local methods). It focused on Bayes modal θ estimates with a N(0,1) prior. A fundamental premise of the global-slope approach is a single linear relationship between θ and $\hat{\theta}$ over the entire θ range. The linearity assumption was found to hold for longer tests (e.g., 30 or more items). As expected, shorter tests (e.g., 15 or fewer items) yielded nearly linear, but slightly S-shaped scatterplots of θ and $\hat{\theta}_{\text{BME-N(0,1)}}$. Unfortunately, due to a confounding factor caused by the pool, the study was unable to assess the impact of the minor violation of the assumption on the information functions of the shorter tests.

The confounding element was a lack of more difficult items in the pool. How did this pool characteristic affect information functions for the global-slope procedure? First, it caused $\hat{\theta}_{\text{BME-N(0,1)}}$ for high ability simulees to shrink toward the prior mean of 0 and fall short of 3.0, the maximum θ value, on the 10- and 15-item tests. This translated to increasingly smaller conditional SEMs on the higher end when they typically get larger toward the maximum (and minimum) θ values. As noted above, the information function in general has the SEM in the denominator and the slope in the numerator. Because, in the case of the global-slope method, the SEM (the denominator) is "local" (i.e., for one θ value) while the slope (the numerator) is "global" (i.e., constant across the θ range), increasingly smaller SEMs meant increasingly larger information values, hence a fish-hook rise on the upper end for the global-slope method.

Why were information functions for the local and quasi-local methods not influenced by the same pool characteristic, even on the shorter tests? Simply stated, it is because both the slope and the SEM are "local" in the local and quasi-local methods. Take Figure 2 as an example (any *N_k*): The conditional SEM is very small at a θ of, say, 2.8, but the "local" slope must be even smaller— a best-fitting linear line through a tiny area around the θ should be practically flat, that is, near zero. Thus, information on the upper end would be smaller than it is in the middle and unaffected by the smaller SEMs.

Information functions of the local and quasi-local methods that are not influenced by the difficulty of a pool are a mixed blessing. On the one hand, they seem to be more robust in the sense that they produced normal-looking information functions despite the pool characteristics. On the other hand, if you would like to look at score information functions to see if you have a good quality pool, you might find some use for a method that is not as robust and might be more "diagnostic," indicating an area of improvement.

If it is too challenging to improve a pool that is not well balanced in terms of difficulty, there seem to be at least two ways to compensate for the need for more difficult (or easy) items. One way is to use another type of IRT score— the maximum likelihood estimator (MLE). Our previous study (Ito, Pommerich, & Segall, 2009) made the same comparisons as in this study, employing the same pool and simulation conditions. The previous study, however, used the MLE, and the results exhibited nothing unusual. This is not surprising, considering that the MLE is the BME with a uniform prior. There was no shrinkage toward the mean of a prior and therefore no "increasingly smaller SEMs" on the high θ end. Another way is to increase test length. The current study demonstrated that longer tests (e.g., ≥ 30 items) were not impacted by an insufficiently difficult pool, and their information functions looked reasonable for all three methods and comparable. That said, it probably is not easy to introduce these changes to an already established testing program. CAT testing programs that are just starting might take these points into consideration in deciding on test length and/or the type of score for reporting.

Future Research

Because the insufficiency of difficult items in the pool clearly confounded the results for the shorter tests, we need to revisit $\text{BME}_{N(0,1)}$ information functions with an augmented pool. Judging from the results for the longer tests which do not appear to have been impacted by the pool characteristics, $\text{BME}_{N(0,1)}$ information functions of the global-slope procedure might tend to be more similar to those using the quasi-local procedure than to those based on the local procedure. However, it is more or less speculative at this point. With the MLE, the global-slope approach produced information functions that tend to be more similar to those of the local approach than to those of the quasi-local approach.

Another alternative method of estimating score information for CAT evolved during the previous study. Because different examinees usually take different sets of items in a CAT administration, there can be as many score (or test) information functions as the number of tests. What if the individual information functions are averaged conditionally on θ or $\hat{\theta}$? That is, compute the score information for all examinees at a given θ (or $\hat{\theta}$) level, average the information values at the θ (or $\hat{\theta}$) level and repeat the process for all θ (or $\hat{\theta}$) levels. Might it be a reasonable way to summarize the amount of information contained in all the $\hat{\theta}$s that are based on different tests? This method (referred to here as the *conditional averaging* method) would be a departure from the typical way of estimating score information as a function of the slope of $\hat{\theta}$ on θ and the standard error of measurement of $\hat{\theta}$ for a given θ, but it might be an interesting comparison.

References

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

Defense Manpower Data Center [DMDC] (2009). *ASVAB Technical Bulletin No. 2*. Seaside, CA: Author.

DMDC (2008). *ASVAB Technical Bulletin No. 3*. Seaside, CA: Author.

du Toit, M. (Ed.) (2003). IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact. Lincolnwood, IL: Scientific Software International.

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

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

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

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

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: Military Testing Association.