The Journal of Experimental Education

Routledge
Taylor & Francis Group

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: http://www.tandfonline.com/loi/vjxe20

Minimizing the Influence of Item Parameter Estimation Errors in Test Development: A Comparison of Three Selection Procedures

Mark J. Gierl , Dianne Henderson , Michael Jodoin & Don Klinger

To cite this article: Mark J. Gierl , Dianne Henderson , Michael Jodoin & Don Klinger (2001) Minimizing the Influence of Item Parameter Estimation Errors in Test Development: A Comparison of Three Selection Procedures, The Journal of Experimental Education, 69:3, 261-279, DOI: 10.1080/00220970109599488

To link to this article: http://dx.doi.org/10.1080/00220970109599488

Published online: 01 Apr 2010.

The Journal of Experimental Education, *2001*, *69*(3), 261–279

Minimizing the Influence of Item Parameter Estimation Errors in Test Development: A Comparison of Three Selection Procedures

MARK J. GIERL
DIANNE HENDERSON
MICHAEL JODOIN
DON KLINGER
University of Alberta

ABSTRACT. In test development, item response theory (IRT) is a method to determine the amount of information that each item (i.e., item information function) and combination of items (i.e., test information function) provide in the estimation of an examinee's ability. Studies investigating the effects of item parameter estimation errors over a range of ability have demonstrated an overestimation of information when the most discriminating items are selected (i.e., item selection based on maximum information). In the present study, the authors examined the influence of item parameter estimation errors across 3 item selection methods—maximum no target, maximum target, and theta maximum—using the 2- and 3-parameter logistic IRT models. Tests created with the maximum no target and maximum target item selection procedures consistently overestimated the test information function. Conversely, tests created using the theta maximum item selection procedure yielded more consistent estimates of the test information function and, at times, underestimated the test information function. Implications for test development are discussed.

Key words: item analysis, item response theory, test development

ITEM RESPONSE THEORY (IRT) provides an appealing conceptual framework for test development (Green, Yen, & Burket, 1989; Hambleton, 1989; Lord, 1980) in large part because of the item information function [i.e., $I_i(\theta)$], which gives a measure of how much information a test item provides at a given ability level. Statistically defined, the item information function is inversely proportional to the square of the width of the asymptotic confidence interval for θ. This relationship implies that the larger the information function, the smaller the confidence interval and the more accurate the measurement. For the two-parameter logistic IRT model, the item information function for item i with ability theta θ is calculated as (Hambleton, Swaminathan, & Rogers, 1991)

$$I_i(\theta) = \frac{D^2 a_i^2}{[e^{Da_i(\theta - b_i)}][1 + e^{-Da_i(\theta - b_i)}]^2},$$

where D equals 1.7, a_i is the item discrimination parameter, and b_i is the item difficulty parameter. For any given ability level, the amount of information increases with larger values of a_i. For the three-parameter logistic IRT model, the item information function is calculated as (Lord, 1980, p. 73)

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{[c_i + e^{Da_i(\theta - b_i)}][1 + e^{-Da_i(\theta - b_i)}]^2},$$

where D equals 1.7, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-chance parameter. For any given ability level, the amount of information increases with larger values of a_i and decreases with larger values of c_i. That is, item discrimination reflects the amount of information an item provides, assuming the pseudo-chance level is relatively small.

The test information function [i.e., $I(\theta)$] is an extension of the item information function. The test information function is simply the sum of the item information functions at a given ability level:

$$I(\theta) = \sum_{i-1}^{n} I_i(\theta),$$

where $I_i(\theta)$ is the item information and n is the number of test items. It defines the relationship between ability and the information provided by a test. The more information each item contributes, the higher the test information function.

Both the item and the test information functions are used in test development. Lord (1980, p. 72) outlined the following four-step procedure for designing a test using calibrated items from a bank:

Step 1: Decide on the shape desired for the test information function. The desire curve is called the target information curve.

Members of this research project contributed equally to the study and to the preparation of the artide. The authors are listed in alphabetical order. We would like to thank the executive editor and the reviewers for their constructive comments on an earlier version of this article. A previous version of this article was presented at the 1999 annual meeting of the American Educational Research Association in Montreal, Quebec, Canada.

Address correspondence to Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, 6–102 Education Centre North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5. E-mail: mark.gierl@ualberta.ca

Step 2: Select items with item information curves that will fill the hard-to-fill areas under the target information curve.

Step 3: Cumulatively add up the item information curves, obtaining at all times the information curve for the part-test composed of items already selected.

Step 4: Continue until the area under the target information curve is filled up to a satisfactory approximation.

One can fill the target information curve using item selection based on maximum information or theta maximum. The first procedure, ***maximum information,*** yields the maximum value of information for an item regardless of its location on the theta scale. For the two-parameter logistic IRT model, maximum information is given by $0.722a_i^2$ (Lord, 1980, p. 151); that is, it reflects item discrimination. For the three-parameter logistic IRT model, maximum information is calculated as (Lord, 1980, p. 152)

$$I_i(\theta)_{\mathrm{MAX}} = \frac{D^2 a_i^2}{8(1-c_i)^2}\left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{3/2}\right],$$

where $I_i(\theta)_{\mathrm{MAX}}$ is the maximum information provided by an item, D equals 1.7, a_i is the discrimination parameter, and c_i is the pseudo-chance parameter. Maximum information is commonly used in test development because it provides a method for selecting the most discriminating items.

Theta maximum is an alternative item selection procedure for filling the target information curve. This method, although less commonly used in test development, provides the location on the theta scale at which an item has the most information. In other words, it determines the peak or the top of the item information function and specifies this location on theta. For the two-parameter model, theta maximum is given by the *b*-parameter for the item (Lord, 1980, p. 152); that is, it reflects item difficulty. For the three-parameter model, theta maximum is calculated as (Lord, 1980, p. 152)

$$I_i(\theta)_{\mathrm{THETA}} = b_i + \frac{1}{Da_i}\,\mathrm{Ln}\,\frac{1+\sqrt{1+8c_i}}{2},$$

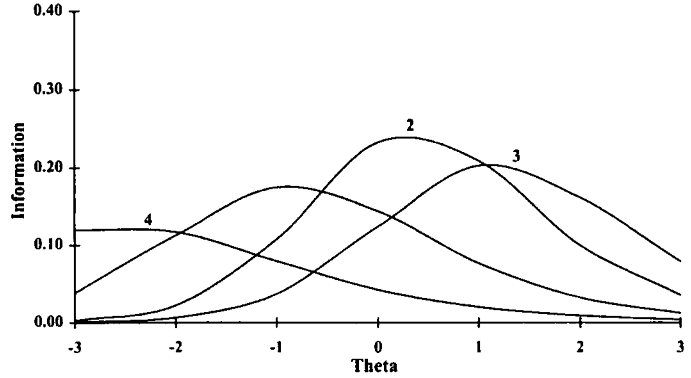
where $I_i(\theta)_{\mathrm{THETA}}$ is the location on the ability scale at which the item information function is maximum, D equals 1.7, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the pseudo-chance parameter.

To illustrate the differences between $I_i(\theta)_{\mathrm{MAX}}$ and $I_i(\theta)_{\mathrm{THETA}}$, we provide an example. Table 1 contains the *a*-, *b*-, and *c*-parameter values for four items, along with the $I_i(\theta)_{\mathrm{MAX}}$ and $I_i(\theta)_{\mathrm{THETA}}$ values for each item. Figure 1 shows the information function for each item. If, on the one hand, the objective in creating a test was to select the most discriminating item from this set, then Item 2 would be chosen because it is the item with the maximum information [i.e., $I_i(\theta)_{\mathrm{MAX}}$ = 0.25]. If, on the other hand, the objective was to select the item that was most discriminating at $\theta = -1.0$, then Item 1 would be chosen because it is the item with

TABLE 1
Item Parameter, Maximum Information, and Theta Maximum Values for Four Example Items

Item	*a*-parameter	*b*-parameter	*c*-parameter	$I_i(\theta)_{\text{MAX}}$	$I_i(\theta)_{\text{THETA}}$
1	0.60	-1.10	0.20	0.18	-1.01
2	0.70	0.14	0.19	0.25	0.25
3	0.64	0.91	0.19	0.20	1.01
4	0.50	-2.87	0.19	0.13	-2.79

FIGURE 1. Information function for four example items.



the maximum information at this point on the theta scale [i.e., $I_i(\theta)_{\text{THETA}} = -1.01$]. Notice that Item 1 is not the item with the most information in this set, but it does yield the most information at $\theta = -1.0$ relative to the other three items. In a comparison of these selection methods for the three-parameter logistic model, maximum information reflects the height of the information curve, which is primarily influenced by the *a*- and *c*-parameters, whereas theta maximum reflects the location of the top of the information function (scaled onto theta), which is primarily influenced by the *b*-parameter. This relationship is more apparent in the two-parameter case, in which maximum information is due solely to the *a*-parameter and theta maximum to the *b*-parameter. In short, each method uses a different criterion for item selection.

Influence of Item Parameter Estimation Errors in Test Development

The item parameters in IRT are merely estimates of true values. They contain estimation errors that result in a correlation between the estimated and true parameters that is less than 1.0. Estimation errors become problematic when one is constructing a test to fit the target information curve, especially when selecting the most discriminating items—that is, items selected using maximum information (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993). Although the most discriminating items will produce a test with maximum information, items with high discrimination also tend to be overestimated relative to their true values, resulting in a test that provides less information than expected (i.e., the actual information curve is less than the target information curve). This outcome can lead to overconfidence in the accuracy of the examinees' ability estimates.

Hambleton et al. (1993) first addressed this problem with a simulation study. Using a two-parameter IRT model, they studied the effects of item bank size, test length, and sample size used in item calibration by comparing the test information function produced by using the estimated parameters with that produced by using the true item parameters. Their major finding was clear: Selecting the most discriminating items from a bank resulted in an overestimated test information function. Moreover, larger banks, shorter tests, and smaller samples used in item calibration resulted in greater estimation error because of the imperfect relation between the estimated and the true parameters.

Hambleton and Jones (1994) extended this research to examine the effects of item parameter estimation errors using a three-parameter logistic IRT model, again with a simulation study, but with realistic item parameters from an actual test administration. Similar to Hambleton et al. (1993), Hambleton and Jones found that the test information function was overestimated when the most discriminating items were selected. The amount of overestimation was influenced by the ratio of item bank size to test length and by the sample size used in item calibration. Large banks resulted in greater estimation errors than did test length and smaller samples used in item calibration. Although these two studies investigated the effects of item parameter estimation errors in test development, neither Hambleton et al. nor Hambleton and Jones included both the two- and the three-parameter IRT models in their study, and both used a single item selection method, maximum information.

We designed this study to replicate and extend the findings reported by Hambleton et al. (1993) and Hambleton and Jones (1994). Our purpose was to compare the $I_i(\theta)_{\text{MAX}}$ and $I_i(\theta)_{\text{THETA}}$ item selection procedures using the two- and three-parameter IRT models. Research suggests that the *b*-parameter is calibrated more accurately than the *a*-parameter, thereby producing a more stable estimate for item selection (Yen, 1987; Yoes, 1996). Moreover, $I_i(\theta)_{\text{THETA}}$ reflects the location rather than the height of the information function. Therefore, one would expect that estimation errors in this function would be reduced when one uses this procedure to select items for a test, thereby overcoming a key problem identified by Hambleton et al. and Hambleton and Jones when IRT is used for test development, namely, the inaccurate estimation of item parameters.

We evaluated this hypothesis using both the two- and three-parameter models. The two-parameter model serves as a comparison to Hambleton et al. (1993), with the addition of the two item selection methods, $I_i(\theta)_{\mathrm{MAX}}$ and $I_i(\theta)_{\mathrm{THETA}}$. We included the three-parameter model because it is generally favored in test development because it provides better fit to multiple-choice data in many testing situations (Hambleton et al., 1991). Therefore, it is important to study both models across selection procedures to evaluate the impact of item parameter estimation errors on the test information function. The results from the current study will provide practitioners with a better understanding of how item parameter estimation errors influence test development when the two- and three-parameter models are used with the $I_i(\theta)_{\mathrm{MAX}}$ and $I_i(\theta)_{\mathrm{THETA}}$ item selection procedures.

Method

To complete the analysis, we simulated three separate banks of items, one for each of the following three conditions: The first bank was generated from the two-parameter logistic model, as in Hambleton et al. (1993). We generated the second bank from the three-parameter logistic model using restricted item characteristics. We generated the third bank from the three-parameter logistic model using realistic item characteristics. We created two conditions for the three-parameter model to compare the results from an efficient item bank typical in a computerized adaptive testing situation (e.g., Flaugher, 1990, p. 46) with those from a realistic item bank typical in a large-scale norm-referenced achievement testing program (i.e., items in the bank, for the most part, provide good discrimination uniformly across the ability score scale). Each bank contained 150 items, and each bank was created with the computer program IRTDATA (Johanson, 1992).

For the two-parameter bank, item discrimination (*M* = 1.00, *SD* = 0.00) and item difficulty values (*M* = 0.00, *SD* = 1.00) were uniformly distributed. These characteristics were also used by Hambleton et al. (1993) for their two-parameter bank. For the three-parameter bank with restricted characteristics, item discrimination (*M* = 1.00, *SD* = 0.00), item difficulty (*M* = 0.00, *SD* = 1.00), and pseudo-chance values (*M* = 0.20, *SD* = 0.00) were uniformly distributed. For the three-parameter bank with realistic characteristics, item discrimination (*M* = 1.00, *SD* = 0.20), item difficulty (*M* = 0.00, *SD* = 1.00), and pseudo-chance values (*M* = 0.20, *SD* = 0.05) were also uniformly distributed. Two randomly equivalent samples, A and B, were drawn from a normal ability distribution (*M* = 0.00, *SD* = 1.00).

For each sample, the simulated item response vectors were generated with samples of size 400, 1,000, and 2,000. The sample size ranged from 400 examinees, a number below the minimum sample size recommended for use with the three-parameter model (Hulin, Lissak, & Drasgow, 1982), to 2,000 examinees, a number above the sample size deemed necessary for accurate parameter estimates with the three-parameter model. These values are also similar to the sample sizes used by Hambleton et al. (1993) and Hambleton and Jones (1994), thereby allowing for a comparison of the results between these former studies and the current study.

Next, we estimated parameters for both Samples A and B using the simulated item response vectors with BILOG 3.09. The default settings in BILOG were used with the exception of the calibration option that was set to "float," indicating that the means of the priors on the item parameters were estimated using marginal maximum likelihood estimation along with the item parameters, and both the means and the item parameters were updated after each iteration (Mislevy & Bock, 1991, pp. 4–27). We used this option because it should result in more accurate item parameter estimation. Because BILOG arbitrarily centers the item parameters and the ability estimates for each calibration, we placed the estimated *a*- and *b*-parameters for Sample B on the same scale as Sample A using the linear equating formula:

$$l_X(y) = x = s(X)\left[\frac{y - \bar{x}(Y)}{s(Y)}\right] + \bar{x}(X),$$

where $l_X(y)$ is the transformed parameter estimate of Sample B placed onto the scale of Sample A and $[\bar{x}(Y), \bar{x}(X)]$ and $[s(Y), s(X)]$ are the means and standard deviations of the Sample B and Sample A parameter estimates, respectively (Kolen & Brennan, 1996).

Using the scaled item parameter estimates, we constructed tests in three ways. First, we created a test by selecting the 25 items from the Sample A bank with maximum information. We chose 25 items to maintain a bank-to-test ratio of 6:1. This ratio can often be found in testing programs that use item banks and was also used by Hambleton et al. (1993). In this condition, no target information curve was specified and the 25 items with the largest information values were selected. This condition is referred to as *maximum no target.*

Second, we created a test by selecting 25 items from the Sample A bank with maximum information at points along a target information curve. In this condition, we evaluated the maximum information procedure in a realistic context using a 25-item test designed to fit a specific target that differentiated examinees throughout the range –2.00 to 2.00. This approach would be appropriate for norm-referenced test construction. Items were selected on the basis of maximum information following a procedure that could be used to fill the target information curve in an actual testing situation. Items with maximum information closest to $\theta = -2.00$ were selected, and this selection procedure continued in increments of 0.50 to the point 2.00. At each theta value, three items with maximum information were selected, except at the two extreme intervals at which only two items were chosen. This condition is called *maximum target.*

Third, we created a test by choosing items from the Sample A bank with the theta maximum item selection procedure. We specified the same norm-referenced target information curve as in the maximum target case to select items throughout the range –2.00 to 2.00. Items with maximum information nearest $\theta = -2.00$ were selected, and this selection procedure continued in increments of 0.50 to the point 2.00. At each theta value, three items were selected, except at the two extremes at which only two items were chosen. This condition is called *theta maximum.*

We assessed estimation errors for the 25-item tests by comparing the test information functions and their relative efficiency of Sample A to those of Sample B. This comparison yielded results that are both meaningful and practical because it allowed us to highlight the effects that chance may produce in a real testing situation (i.e., comparing two estimates rather than comparing an estimate against truth, because truth is never known to practitioners; Hambleton & Jones, 1994, pp. 177–178). A graph of the test information functions across the selection procedures demonstrates how these methods compare. When the information for Sample A exceeded the information for Sample B, the item parameters contained estimation errors. In addition, one can compare the total amount of information across the three selection procedures by examining the height of the information function for each test. Relative efficiency, on the other hand, demonstrates how each selection procedure compares with itself across validation samples by relating the Sample B test, which served as the baseline, to the Sample A test, which served as the cross-validation condition. When the cross-validation condition is more efficient than the baseline, the item parameters contain estimation errors.

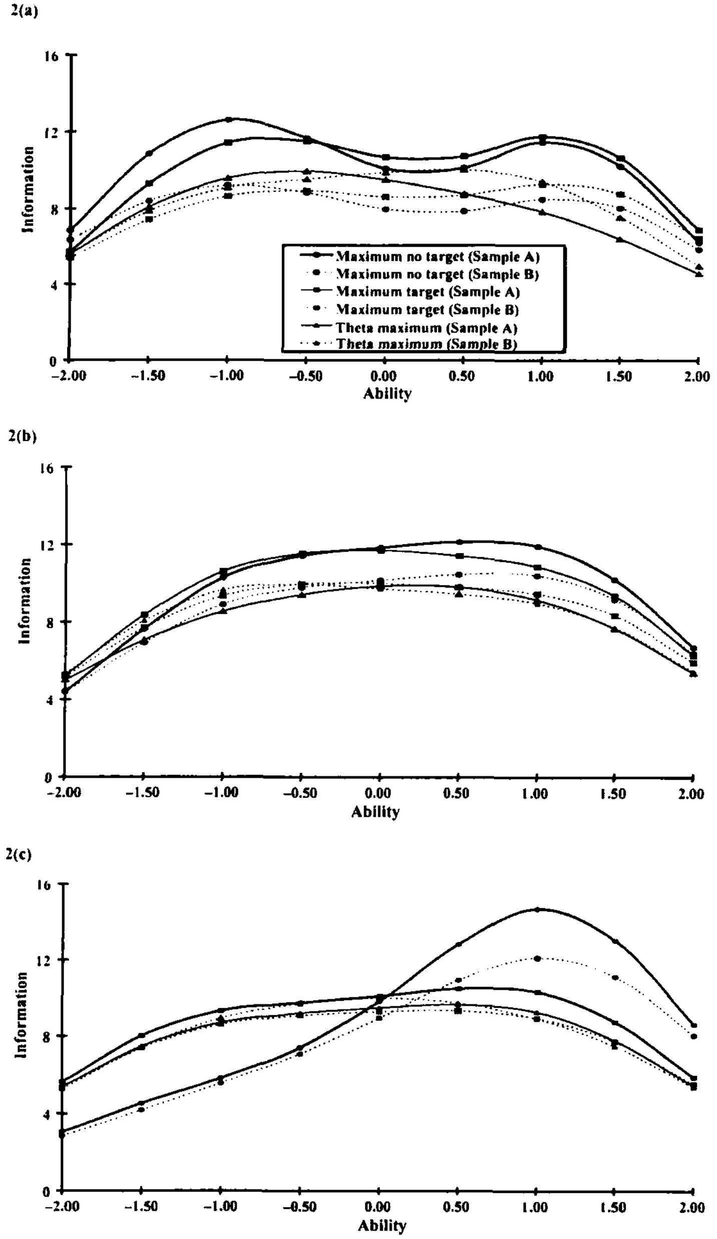
Results

Two-Parameter Model

Figure 2 illustrates the item parameter estimation errors, as manifest in the test information functions, associated with the selection methods using Sample A items from the two-parameter bank for samples of 400, 1,000, and 2,000 simulees, respectively. As expected, items selected with maximum no target and maximum target resulted in an overestimation of the test information function when Samples A and B were compared. This effect was less pronounced as sample size increased. That is, the test information functions for Samples A and B were more comparable as the number of simulees increased from 400 to 2,000. This outcome is consistent with the finding reported by Hambleton et al. (1993), namely, that the item parameters from the maximum information (no target) condition contain estimation errors that result in overestimation of the target information function.

As sample size increased, the test information functions for maximum no target, maximum target, and theta maximum became somewhat platykurtic and

FIGURE 2. Comparison of the information functions for the two-parameter logistic IRT model using three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.



approximately uniform in the range –1.50 to 1.50. The one exception occurred with 2,000 simulees in the maximum no target condition in which the functions were more leptokurtic compared with the functions from the other two tests. Maximum no target produced a 25-item test with more information compared with maximum target and theta maximum, but only in the ability range above 0 on the theta scale. Maximum target also produced a test with more information compared with the theta maximum item selection procedure for ability estimates greater than 0.

The relative efficiency of each test across selection procedures at sample sizes of 400, 1,000, and 2,000, respectively, is presented in Figure 3. In these comparisons, Sample B was fixed at 1, with overestimates of information occurring for values greater than 1 and underestimates for values less than 1. These graphs clearly demonstrate that the maximum no target and maximum target procedures produced inaccurate information estimates when the cross-validation and baseline samples were compared. The amount of information estimation error decreased as sample size increased. Alternatively, the relative efficiency for the cross-validation sample with the theta maximum method was much closer to that for the baseline sample, with over- and underestimation of the information function occurring at various points along the theta scale. This trend is apparent across all three sample sizes, indicating that theta maximum tended to have smaller information estimation errors than the maximum no target and maximum target procedures when the two-parameter model was used.

Three-Parameter Model—Restricted Item Parameters

Figure 4 illustrates the information estimation errors associated with the three selection procedures using Sample A items from the restricted three-parameter bank. Again, maximum no target and maximum target consistently overestimated—whereas theta maximum both over- and underestimated—the test information function. This outcome occurred in all three sample sizes.

The test information functions for maximum no target were leptokurtic and negatively skewed, whereas the test information functions for maximum target and theta maximum were more platykurtic and uniform. Maximum no target produced a 25-item test with more information compared with maximum target and theta maximum, but only in the ability range above 0 on the theta scale. Maximum target also produced a test with more information than theta maximum, but these differences became small as sample size increased.

The relative efficiency of each test across selection procedures and samples is presented in Figure 5. Maximum no target and maximum target tended to overestimate information, whereas theta maximum both over- and underestimated the function. Differences between the cross-validation and baseline samples were much smaller with theta maximum compared with the maximum no target and

FIGURE 3. Relative efficiency for the tests created with the two-parameter logistic IRT model across three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.

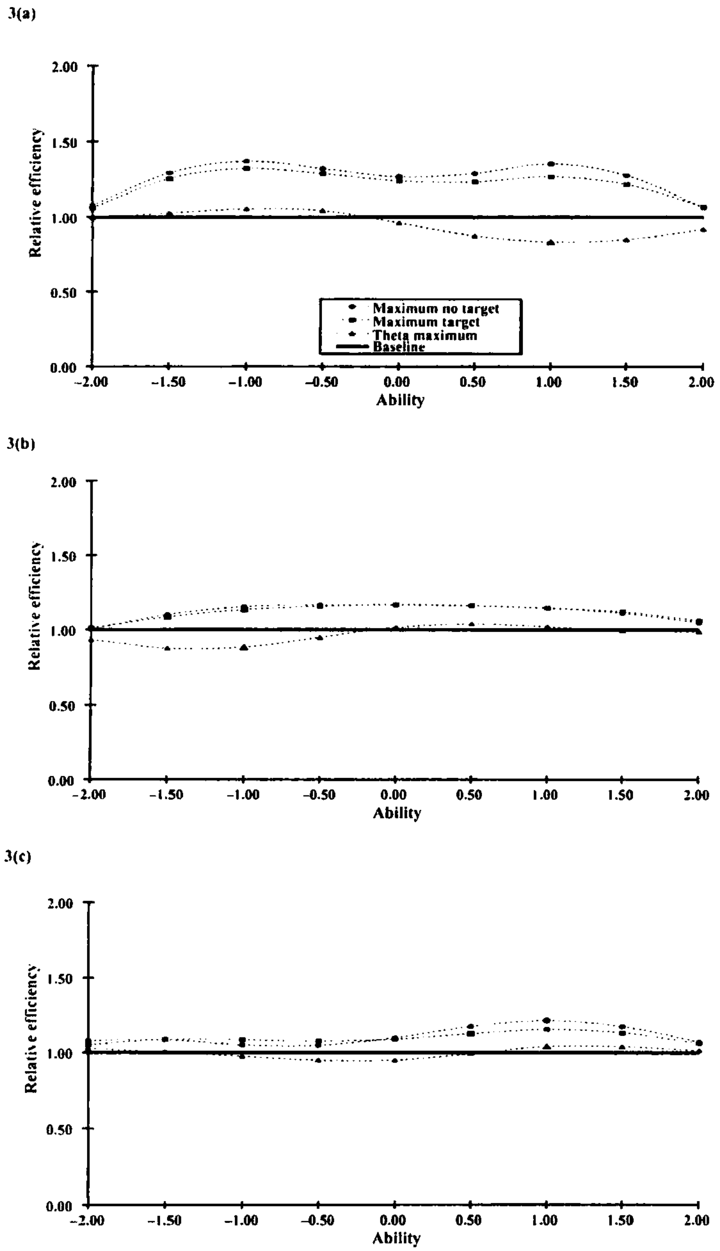


FIGURE 4. Comparison of the information functions for the restricted three-parameter logistic IRT model using three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.

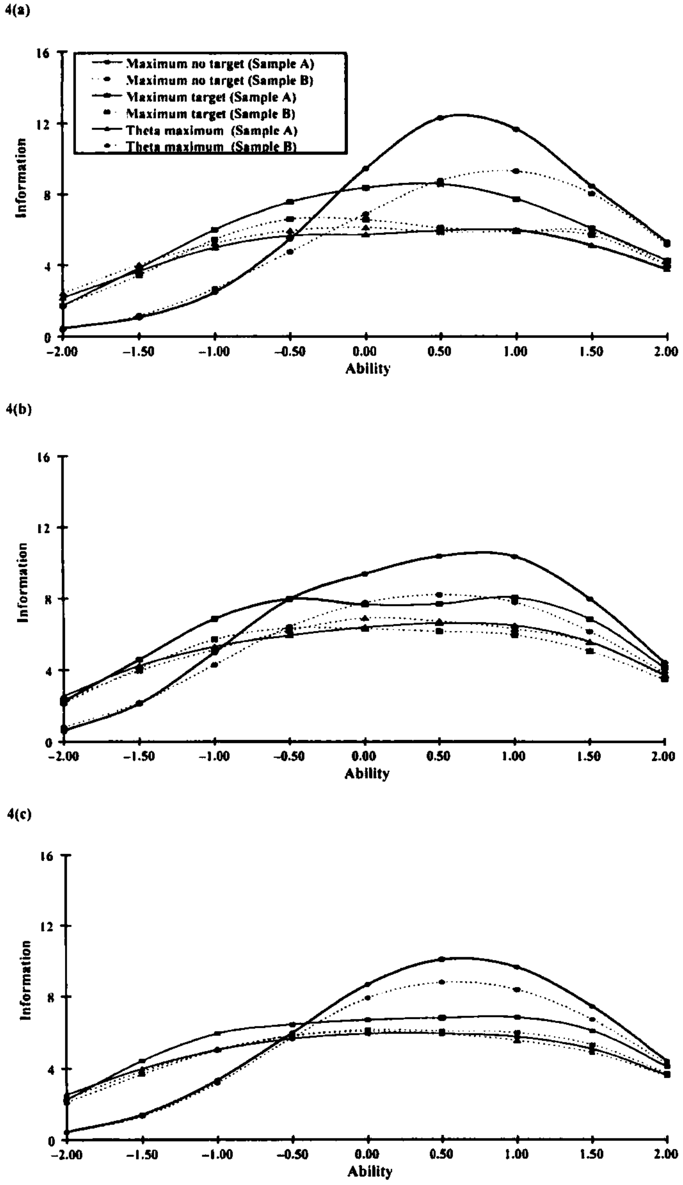


FIGURE 5. Relative efficiency for the tests created with the restricted three-parameter logistic IRT model across three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.

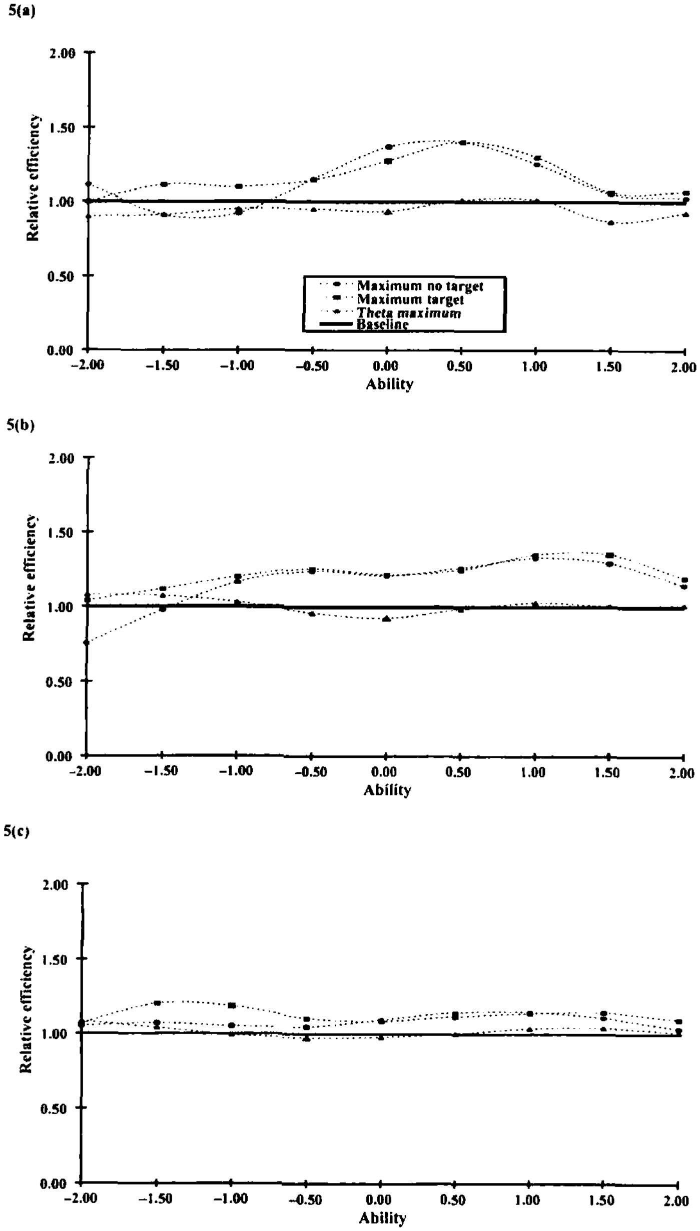


FIGURE 6. Comparison of the information functions for the realistic three-parameter logistic IRT model using three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.

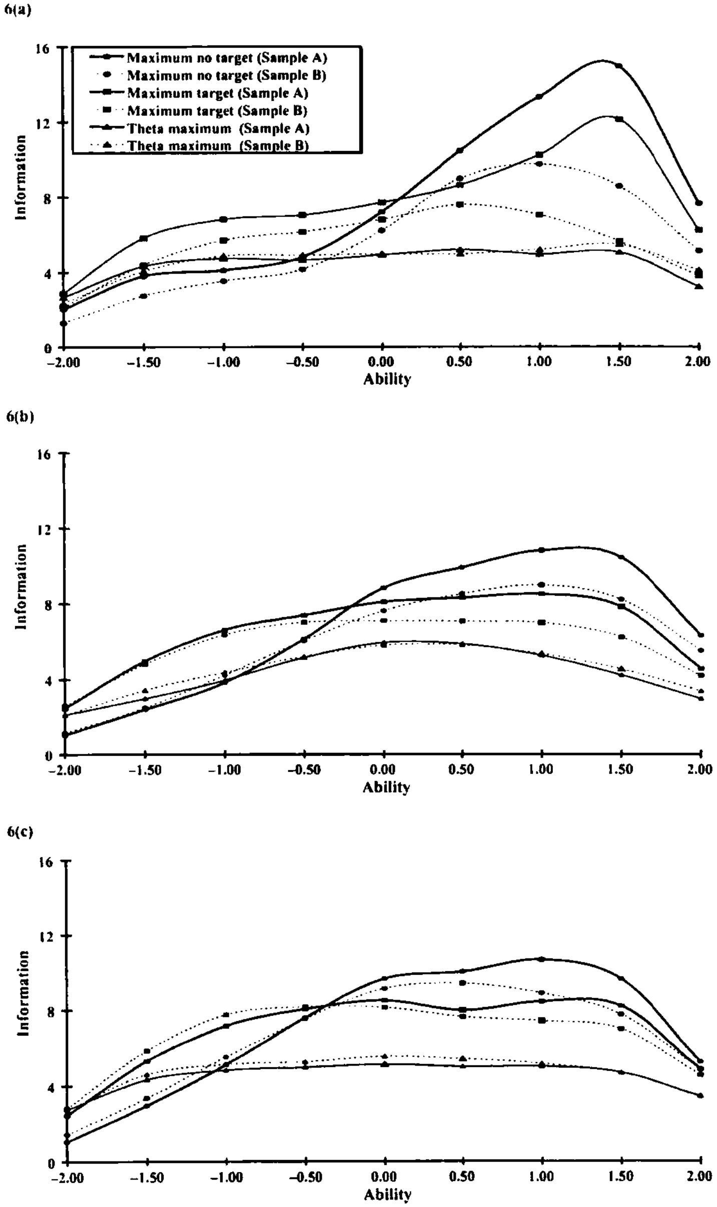
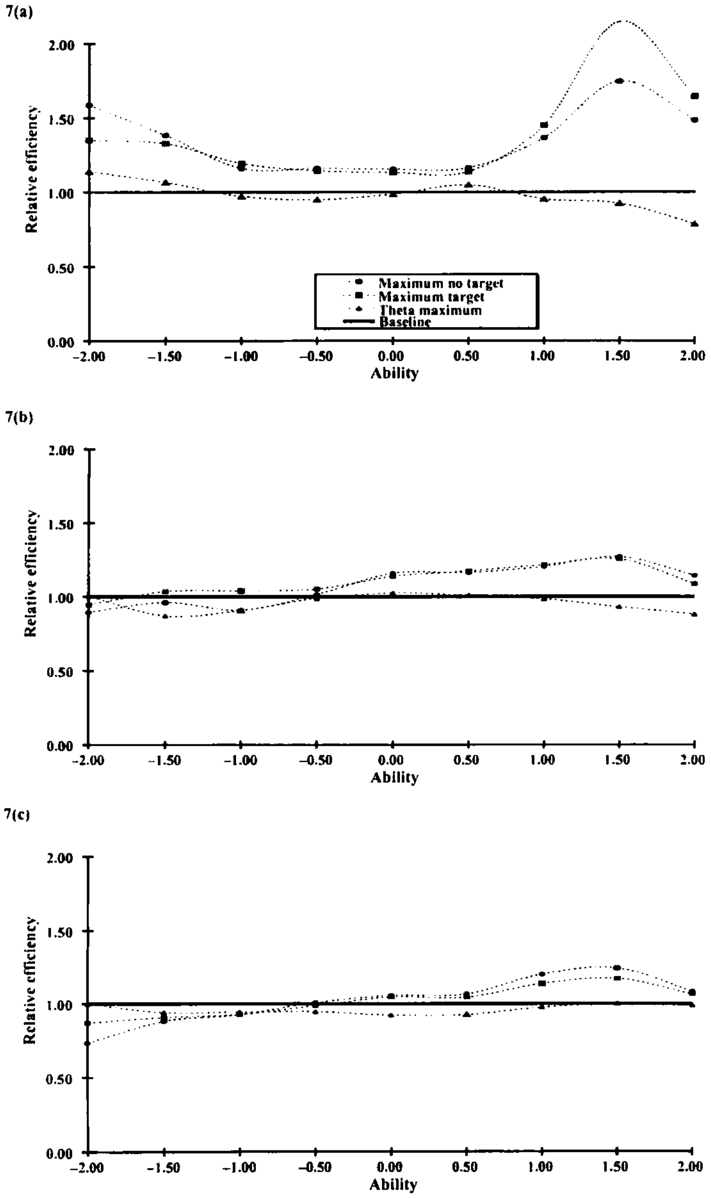


FIGURE 7. Relative efficiency for the tests created with the realistic three-parameter logistic IRT model across three item selection methods with (a) 400, (b) 1,000, and (c) 2,000 simulees.



maximum target procedures, although these differences decreased in all three selection procedures as sample size increased.

Three-Parameter Model—Realistic Item Parameters

Figure 6 illustrates the item parameter estimation errors associated with the three selection procedures with Sample A items from the realistic three-parameter bank. Again, maximum no target and maximum target consistently overestimated, whereas theta maximum over- and underestimated the information function across Samples A and B. This outcome is consistent with the Hambleton and Jones (1994) finding that item selection based on maximum information (no target) results in an overestimation of the test information function.

The test information function for maximum no target was leptokurtic and negatively skewed, whereas the functions for maximum target and theta maximum were platykurtic and relatively uniform, especially at the larger sample sizes. Maximum no target produced a 25-item test with more information compared with maximum target and theta maximum, but only in the ability range above 0 on the theta scale. Moreover, unlike the result from the restricted three-parameter condition, maximum target yielded a test with more information than theta maximum, and that difference remained relatively constant across the 400-, 1,000-, and 2,000-simulee conditions.

The relative efficiency for each test is presented in Figure 7. Maximum no target and target, once again, tended to overestimate the information function, whereas theta maximum was more consistent when the cross-validation and baseline samples were compared with both over- and underestimates. The influence of sample size was most apparent in this three-parameter condition because overestimation of the information function for the maximum no target and maximum target procedures was larger in the 400-simulee condition than in the 1,000- and 2,000-simulee conditions, in which the amount of item parameter estimation error was much smaller.

Discussion

The purpose of this study was to compare three item selection procedures with two IRT models commonly used in test development to better understand the influence of item parameter estimation errors. These errors are problematic when one is developing a test, especially when the most discriminating items are selected (i.e., items with maximum information), because these items tend to be overestimated relative to their true values, resulting in a test that provides less information than expected. This outcome is problematic because it can lead to overconfidence in the accuracy of examinees' ability estimates. In a comparison of the item selection procedures, maximum no target and maximum target reflect the height of the information curve that is primarily influenced by the *a*-parameter for the two-parameter model and the *a*- and *c*-parameters for the three-parameter model. Alternatively, theta maximum identifies the top of the information function and scales this location onto theta. Theta maximum is primarily influenced by the *b*-parameter for both the two- and three-parameter models. Because theta maximum reflects the location (i.e., *b*-parameter) rather than the height (i.e., *a*-parameter) of the information curve, we expected that errors associated with the test information function would be reduced when this item selection procedure was used.

Differences among the item selection methods were evident. For the two-parameter model, tests created with maximum no target, maximum target, and theta maximum produced similar test information functions with samples of 400 and 1,000 simulees, although the maximum no target and maximum target procedures consistently produced tests with more information than the theta maximum procedure. With 2,000 simulees, maximum no target produced a test with notably higher information above 0 (and lower below 0) on the theta scale, whereas the maximum target and theta maximum procedures were similar to one another. Across all three sample sizes, the maximum no target and maximum target item selection procedures consistently overestimated—whereas theta maximum both over- and underestimated—the information function when the cross-validation and baseline samples were compared. In addition, the magnitude of estimation error was generally smaller for theta maximum.

For the restricted three-parameter model, the tests created with maximum no target generally produced the most information. Conversely, the tests created with maximum target and theta maximum had less information and were more comparable to one another, especially as sample size increased. Tests created with the maximum no target and maximum target procedures had the largest item parameter estimation errors, whereas tests produced with theta maximum had the smallest item parameter estimation errors.

For the realistic three-parameter model, tests created with maximum no target, again, produced the most information. In addition, tests created with maximum target and theta maximum were noticeably different from one another across the 400, 1,000, and 2,000 simulees, with the maximum target tests having more information than the theta maximum tests. This outcome represents a key difference from the result in the restricted three-parameter condition. Tests created with maximum no target and maximum target procedures also tended to have the largest overestimation of information, especially above 0 on the theta scale. Tests created with theta maximum tended to have the smallest item parameter estimation errors when information functions were compared for the cross-validation and baseline samples. One reviewer of this article noted that regression to the mean could help explain the Sample A-to-B differences across test information functions, stating that

> for both the maximum methods [no target and target], when the item discrimination parameter is over-estimated and the pseudo-chance parameter is under-estimated, and items are chosen on the basis of having extreme (high or low, respectively) value, then the likelihood of the regression effect should be obvious.

This explanation is reasonable, and it underscores one of the main findings in this study, namely, that item selection based on maximum no target and maximum target procedures resulted in an overestimation of the test information function. Conversely, item selection based on theta maximum resulted in test information functions that were more consistent across samples with little or no regression to the mean.

The shape of the test information function for the maximum no target three-parameter conditions (i.e., leptokurtic and negatively skewed) is also noteworthy. This shape likely came about because the *c*-parameter was poorly estimated (e.g., Yen, 1987; Yoes, 1996). As a result of this estimation problem, many items at the lower end of the ability scale probably had inflated *c*-parameters, which reduced the amount of available information. In other words, the items of low difficulty were less informative than the items of high difficulty because of the inflated *c*-parameter.

The results from this study highlight three important findings. First, as was demonstrated by Hambleton et al. (1993) and Hambleton and Jones (1994), item selection based on maximum no target produced an overestimate of test information. This result was also found in the present study with the two- and three-parameter models. Maximum target also tended to overestimate the information function. Alternatively, theta maximum provided the most consistent estimates of the information function. This finding suggests that specifying a target curve and selecting items on the basis of theta maximum has a distinct advantage over the maximum no target and maximum target item selection procedures because they result in a more conservative information function estimate, albeit at the expense of total information. In other words, tests created with theta maximum item selection tend to have less information than tests created with the maximum no target and maximum target item selection procedures, but the information estimates are more reliable.

Second, the results were consistent for each IRT model. This finding suggests that item parameter estimation errors associated with the selection procedures hold across different test banks. Researchers must evaluate these outcomes using item banks with more diverse item characteristics (e.g., including parameters from alternative items formats and creating tests with the parameters from these alternative item formats) and using different target information functions (e.g., using a bimodal target information function for a criterion-referenced testing situation that contains two cut-scores on the ability score scale).

Third, sample size is important. Estimation error consistently decreased as sample size increased. These errors were most extreme in the 400-simulee condition with the realistic three-parameter model using the maximum no target and maximum target item selection procedures. Overestimation was less pronounced with the theta maximum procedure. This finding, again, highlights the important tradeoff between using the maximum no target and maximum target item selection procedures, which produce a peaked test information function while also producing a more unreliable outcome, and using the theta maximum item selection procedure, which produces a comparatively less peaked test information function but a more reliable outcome. Regardless, practitioners should draw adequate samples for item calibration when using IRT in test development.

REFERENCES

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 41–63). Hillsdale, NJ: Erlbaum.

Green, R., Yen, W. M., & Burket, G. R. (1989). Experiences in the applications of item response theory in test construction. *Applied Measurement in Education, 2,* 297–312.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: American Council on Education, Macmillan.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7,* 171–186.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30,* 143–155.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item response curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249–260.

Johanson, G. A. (1992). IRTDATA: An interactive or batch PASCAL program for generating logistic item response data [Computer program]. Athens: Ohio University.

Kolen, M. J., & Brennan, R. L. (1996). *Test equating: Methods and practices.* New York: Springer.

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

Mislevy, R. J., & Bock, R. D. (1991). BILOG 3: Item analysis and test scoring with binary logistic test models [Computer program]. Moorseville, IN: Scientific Software.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52,* 275–291.

Yoes, M. E. (1996). *An updated comparison of microcomputer-based item parameter estimation procedures used with the 3-parameter IRT model* (Assessment Systems Corporation Tech. Rep. 95–1R). St. Paul, MN: Assessment Systems Corporation.