LSAC RESEARCH REPORT SERIES

Capitalization on Item Calibration Error in Computer Adaptive Testing

Wim J. van der Linden
Cees A. W. Glas
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 98-04
September 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

In adaptive testing, each subsequent item is often selected to have maximum information at the current estimate of the ability of the test taker. An important advantage of this procedure, compared to nonadaptive assessments, is that the same measurement precision can be realized at a shorter test length. However, because the properties of the items are estimated from previous response data, the adaptive procedure may capitalize on these estimation errors rather than the true properties of the items.

The problem of capitalization on estimation error has been addressed before in the educational measurement literature in the context of assembling a fixed form of a test. However, the problem has never been addressed for a test with an adaptive format. In this paper, the problem is explored for adaptive testing both through an informal analysis and an empirical study using simulated data.

As expected, adaptive procedures are most sensitive to errors in the estimated discrimination parameters. Also, the simulation study showed a clear preference by the adaptive procedures for items with larger errors in the estimates of the discrimination parameters. However, the effect of this capitalization on estimation error in the item parameters was minor compared to the effects of the relative deficiencies in the item pool corresponding to certain intervals of the ability scale.

The practical conclusion from this study is that good item calibration is an important requisite for adaptive testing. However, once the sample of test takers is large enough, the composition of the item pool takes over as a more important factor for the quality of the adaptive procedure.

Abstract

In adaptive testing, item selection is sequentially optimized during the test. Since the optimization takes place over a pool of items calibrated with estimation error, capitalization on these errors is likely to occur. How serious the consequences of this phenomenon are depends not only on the distribution of the estimation errors in the pool or the ratio of the test length to the pool size, but also on the structure of the item-selection criteria used. A simulation study demonstrated the existence of the phenomenon empirically. It also showed that its effect on the errors in the ability estimates interacts strongly with the distribution of the items in the pool.

Introduction

The ideal underlying computerized adaptive testing (CAT) is to adapt the properties of the test items optimally to the ability of the examinee. An effective framework to realize this goal is item response theory (IRT). An important feature of IRT models is that they have separate parameters to represent the properties of the items and the ability of the examinee. As a consequence, these models can be used to select items such that an optimal match is obtained between (a function of) the values of the item parameters and the value of the ability parameter. Since the value of the ability parameter is not known, the test begins with an a priori estimate of the value of the ability parameter that is updated after each new item response. The values of the item parameters are estimated in advance; during the test these estimates are usually treated as if they are the true values of the parameters. A more complete description of adaptive testing is given in Wainer (1990).

One of the functions of the item parameters often used in adaptive testing is Fisher's information function (Hambleton & Swaminathan, 1985, chap. 6; Lord, 1980, chap. 5). This function not only has the advantage of being monotonically related to the (asymptotic) standard error of the ML estimator of the ability parameter but is also additive in the item information functions. Use of the function is generally accompanied by the application of the maximum-information criterion of item selection which selects the next item to have maximum information at the current estimate of the value of the ability parameter. If the value of the ability parameter is estimated in a Bayesian fashion; that is, by its posterior distribution given the responses on the previous items, other functions of the item parameter values are used. A well-known example of these functions is the one used in the minimum expected posterior variance criterion. In Bayesian adaptive testing, the next item is selected to minimize this function. A more complete description of these item-selection criteria is given below.

Application of an item-selection criterion over a pool of items for a given examinee always involves optimization; that is, the next item is chosen to have a maximum or minimum value for the criterion. However, since the values for the item parameters are estimated, a process generally known as "capitalization on chance" may occur. The process operates on the fact that optimal values of a function of the item parameters can be the result of extreme true values of the parameters as well as large estimation errors. Consequently, if items are selected optimizing the value of this function, large estimation errors tend to be overrepresented among the items selected. The result is an ability estimator with an accuracy likely worse than expected.

In test theory, the phenomenon of capitalization on chance has been well addressed for the selection problem of choosing a battery of variables with the largest predictive validity for job performance or academic success. The measure usually taken to counter its effect is to split the sample into a screening sample and a calibration sample. The variables are then selected in the screening sample but their regression parameters are re-estimated in the calibration sample (Lord & Novick, 1968, chap. 13). The effect of this cross validation typically is a shrinkage of the initial estimates of the regression parameters to more realistic sizes.

The problem of capitalization on chance was not addressed in the literature on test assembly until recently in papers by Hambleton and associates (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993). These authors show that if test forms are assembled to have maximum information over an ability interval and the values of the item parameters are estimated from a sample of N = 400 examinees, the height of the information function may be overestimated by as much as 25 to 40%. Samples of this size are not uncommon in educational testing.

Several factors can be expected to have an impact on the process of capitalization on calibration error. The first is the distribution of the errors in the estimated parameter values in the item pool. Obviously, the larger the errors (or the smaller the calibration sample), the larger the effect of the capitalization on the values of the criterion. The second is the ratio of the number of items selected to the number in the pool. The smaller the ratio, the larger the likelihood of selecting items only from those with the larger estimation errors. The roles of both factors were confirmed in the studies by Hambleton et al.

The authors of this paper had no strong prior opinion as to the question of whether the effects of capitalization on error in CAT would be more or less serious than that in the assembly of test forms with a fixed format. The size of the estimation errors and the selection ratio were certainly expected to remain important factors, but the role of the two new factors was unclear. The first new factor is the structure of the function of the item parameters used in the item-selection criterion. As shown in an analysis below, item-selection criteria are certainly sensitive to estimation error. On the other hand, it is known that for CATs of realistic length, the ability estimator is quite robust with respect to the choice of the item-selection criterion (Chang & Ying, 1996; van der Linden, 1998; van der Linden & Reese, 1998; Veerkamp & Berger, 1997). The same may thus hold true with respect to variation in the criterion values due to estimation error. The second factor deals with the question of how the effects of early capitalization on errors in a CAT propagate later on in the test. In another context, it has been found that early bias in the ML ability estimator in a CAT tends to be neutralized by the maximum-information criterion later in the process (van der Linden, 1998). However, not much is known with respect to the effects of errors in the estimated values of the item parameter.

From a practical point of view, errors due to capitalization on chance in CAT are much more serious than in the assembly of forms for paper-and-pencil testing. All items are selected in real time, and the estimated of their parameter values are used immediately to find the next "optimal" item. In adaptive testing, cross validation of item selection is impossible.

The remainder of this paper is organized as follows. First, the item-selection criteria used in this study are introduced and analyzed for their liability to errors in item parameter estimation. Then, the design of the simulation study is discussed. The last section of the paper presents the results from the simulation study and draws some practical conclusions.

Item-Selection Criteria and Estimation Error

As already indicated, the effects of capitalization on calibration error in CAT may depend not only on the size of the calibration errors but also on the function defined on the item parameters optimized. One of the functions in use for CAT is Fisher's information function. For dichotomously scored items, the function has the following form:

$$I_i(\theta) \equiv \frac{P'(\theta)_i^2}{P_i(\theta)Q_i(\theta)}, \tag{1}$$

$P_i(\theta)$ being the response function for item i, $P'(\theta)$ its first derivative with respect to θ, and $Q_i(\theta) \equiv 1 - P_i(\theta)$ (Lord, 1980, sect. 5.4). In CAT, the function is used to find the item in the pool that yields the largest value at $\theta = \hat{\theta}$, where $\hat{\theta}$ is the current estimate of the ability of the examinee.

For the two-parameter logistic (2-PL) model

$$P_i(\theta) \equiv \left\{1 + \exp\left[-a_i\left(\theta - b_i\right)\right]\right\}^{-1}, \tag{2}$$

with a_i and b_i being the discrimination and difficulty parameter of item i, respectively, the information function is equal to

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta). \tag{3}$$

Analytically, for a fixed value of a_i the function in Equation 3 reaches a maximum for $\theta = b_i$, that is, for the θ value that gives $P_i(\theta) = .50$. At this point the maximum is equal to $.25a_i^2$. Thus, a CAT algorithm based on the maximum-information criterion will have a tendency to select items from the pool with values of b_i close to $\hat{\theta}$ and large values for a_i.

The critical factor in Equation 3 is the size of the discrimination parameter a_i rather than the factor $P_i(\theta)Q_i(\theta)$. Because the parameter is squared in Equation 1, the effect of estimation errors is enlarged. On the other hand, the factor $P_i(\theta)Q_i(\theta)$ in Equation 1 is robust with respect to values for b_i in the neighborhood of the θ value of the examinee, even for larger values of a_i. If the value of $P_i(\theta)$ is in the range of [.40, .60], the maximal difference between the product $P_i(\theta)Q_i(\theta)$ and its maximum value is .01. If the range is enlarged to [.30, .70], the difference is still not larger than .04. Thus, a CAT algorithm based on the maximum-information criteria can be expected to capitalize on large errors in a_i but to be relatively robust with respect to errors in b_i.

If the three-parameter logistic (3-PL) model

$$P_i(\theta) \equiv c_i + (1 - c_i)\left\{1 + \exp\left[-a_i(\theta - b_i)\right]\right\}^{-1} \tag{4}$$

with guessing parameter c_i is chosen, the structure of the information function remains identical to the one in Equation 3. The only change is the replacement of the factor $P_i(\theta)Q_i(\theta)$ in Equation 3 by

$$\left[\left\{\frac{P_i(\theta) - c_i}{1 - c_i}\right\}^2 / P_i(\theta)\right] Q_i(\theta), \tag{5}$$

with $P_i(\theta)$ defined by Equation 4 and $Q_i(\theta) \equiv 1 - P_i(\theta)$. Note that Equation 5 is generally smaller than the factor $P_i(\theta)Q_i(\theta)$ in Equation 3 but that equality is obtained if $c_i \rightarrow 0$. It can therefore be concluded that Equation 5 has a smaller effect on the value of the information function than the factor $P_i(\theta)Q_i(\theta)$ in Equation 3 and that the value of the discrimination parameter a_i remains the critical factor.

A Bayesian criterion for item selection in CAT is the one of minimum expected posterior variance. An approximate version of the criterion for use in CAT was introduced by Owen (1975). In the criterion it is assumed that the ability estimation starts with a prior distribution for θ updated after each item response using Bayes theorem. The next item selected has a predicted posterior distribution with minimum variance among all items. For a more detailed description of this criterion, see van der Linden (1998).

To present the criterion more formally, let $(u_1, \ldots, u_{k-1})$ be the responses obtained on the first k – 1 items in the CAT. If item i is selected, the expected posterior variance is

$$\sum_{j=1}^{2} P_i\left(U_i = j \mid u_1, \ldots, u_{k-1}\right) Var\left(\theta \mid u_1, \ldots, u_{k-1}, U_i = j\right), \tag{6}$$

where $Var\left(\theta \mid u_1, \ldots, u_{k-1}\right)$ is the posterior variance of θ and

$$P_i\left(U_i = j \mid u_1, \ldots, u_{k-1}\right) = \int P_i\left(U_i = j \mid \theta\right) g\left(\theta \mid u_1, \ldots, u_{k-1}\right) d\theta \tag{7}$$

is the posterior predictive probability of response u_i on item i given the responses $u_1, \ldots, u_{k-1}$ to the previous items. The next item is selected to have a minimal value for Equation 6 among the items in the pool.

A variation of the criterion in Equation 6 is the maximum expected posterior-weighted information criterion. The criterion also predicts the probabilities of responses $U_i = 1$ and $U_i = 0$ for each item i in the pool, but uses these probabilities to calculate the expected posterior-weighted information.

$$\sum_{j=1}^{2} P_i\left(U_i = j \mid u_1, \ldots, u_{k-1}\right) \int I_{U_1, \ldots, U_{k-1}, U_i}(\theta)\, g\left(\theta \mid u_1, \ldots, u_{k-1}, U_i = j\right) d\theta \tag{8}$$

where $g(\theta \mid .)$ is the posterior density of θ after $k - 1$ items have been selected. The integral in Equation 8 is the information in the item response vector weighted by the posterior uncertainty if item i is selected as the kth item and response $j = 0,1$ is given to the item. The rest of the expression is to take the expectation over the posterior probabilities of giving response $j = 0,1$ to item i.

The critical difference between the maximum-information criterion in Equation 3 and the maximum expected posterior-weighted information in Equation 8 is the role of the posterior distribution of θ. In Equation 3 the information function is evaluated close to the center of the posterior distribution of θ whereas in Equation 8 the information function is integrated over the full posterior. It is expected that the two criteria show different behavior at the beginning of the test where Equation 3 has a preference for information functions that peak at the center of the posterior, but that difference disappears as the posterior itself becomes peaked later in the test.

Simulation Study

To further explore the role of capitalization on error in CAT, a simulation study was conducted. The effects of the following factors were studied:

1. The size of the calibration sample (N = 500, 1,500, 2,500, ∞);
2. The length of the test (n = 10, 20, 40);
3. The size of the item pool (k = 40, 80, 400, 1,200);
4. The nature of the item-selection criterion (maximum information, minimum expected posterior variance, maximum expected posterior-weighted information).

In all cases, ability was estimated using the expected a posteriori (EAP) estimator with a $N(0,1)$ prior. For the maximum information criterion, ability was also estimated using the weighted maximum likelihood (WML) estimator derived in Warm (1989). The latter is attractive because of its negligible bias.

Method

A calibrated pool of items was simulated as follows: A data matrix with 1,000 examinees by 100 items was available from a Dutch national school graduation exam of English as a foreign language. The items were calibrated under the 2-PL model (see Equation 2) using the method of marginal maximum likelihood estimation with a $N(0,1)$ distribution for the ability parameter. In addition, the information matrix for the item parameters was estimated from the data. To simulate calibration samples of different sizes, the required number of examinees were drawn from the data matrix at random and with replacement. As the information matrix is additive in the examinees, it could easily be adapted to the various samples of examinees.

The true parameter values were equated to the values estimated from the data matrix; their distributions are displayed in Table 1. The distribution of the values for the item difficulty parameter had a mean of .970, for an ability distribution with mean and standard deviation normed at .00 and 1.0, respectively. Thus, the item pool was relatively difficult for the examinees.

TABLE 1
Distribution of true parameter values in simulated item pool

	Mean	Minimum	Maximum	Standard Deviation
a_i	0.777	0.222	1.841	0.288
b_i	0.970	–1.262	3.590	0.885

Item calibration errors were drawn from normal distributions using the information matrix to calculate their variances. To simulate calibrated pools with larger numbers of items, the set of true values of the item parameters was duplicated and independent draws for the error distributions were made.

Each of the item pools in this study had 1,200 simulated items. In one part of the study, the item pool consisted of a mixture of items calibrated using different sample sizes; one quarter of the items were simulated to be calibrated on a sample of 250 examinees, one third on a sample of 1,500 examinees, one third on a sample of 2,500 examinees. These sections of the pool thus had identical distributions of their true parameter values but differed in the size of their calibration errors. The presence of capitalization on calibration errors was examined by counting the number of times items from the three sections were used in the adaptive tests.

In the second part of the study, the item pools were homogeneous with respect to the size of the calibration sample. These pools were used to access the effect of item calibration error on the final ability estimates in the adaptive procedures.

The adaptive testing procedure was replicated 100 times for θ = 2.0, –1.0, 0.0, 1.0, 2.0, to obtain stable estimates of the counts and mean absolute errors.

Results

Figures 1–3 display the counts of the numbers selected in the adaptive procedure from the sections in the item pool calibrated on samples of N = 500, 1,500, and 2,500 examinees as a function of θ. In each panel, the curves always sum to 100n (that is, the number of replications times test length). The dominant impression from the figures is that the smaller the calibration sample size, the larger the number of items selected. A surprisingly strong effect was present for the maximum posterior-weighted expected information criterion in combination with tests of n =10 items. However, an exception was obtained for the maximum-information criterion and WML ability estimation for n = 10; an explanation for this anomaly could not be found. The effect showed a tendency to decline for tests with 40 items but was still present at this test length, in

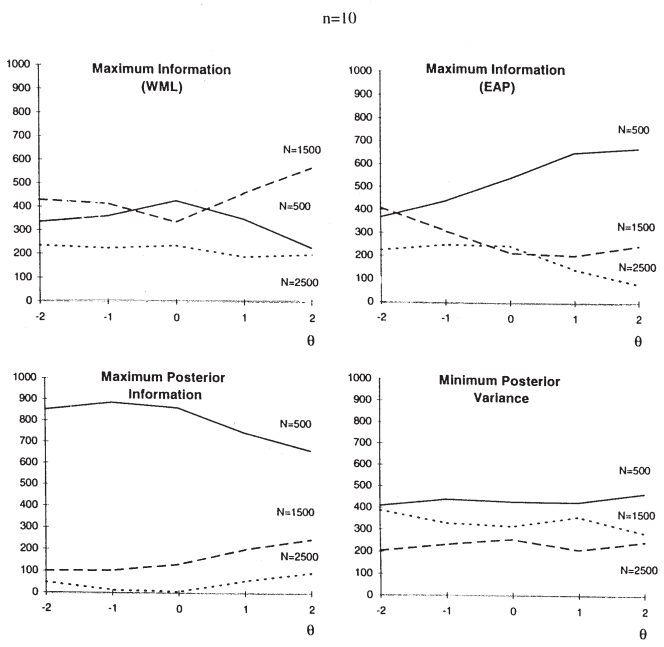

FIGURE 1. *Number of items in the adaptive tests selected from the sections in the pool calibrated on N = 500, 1,500, and 2,500 examinees for the various item-selection criteria (n = 10)*

particular at the high end of the ability scale.

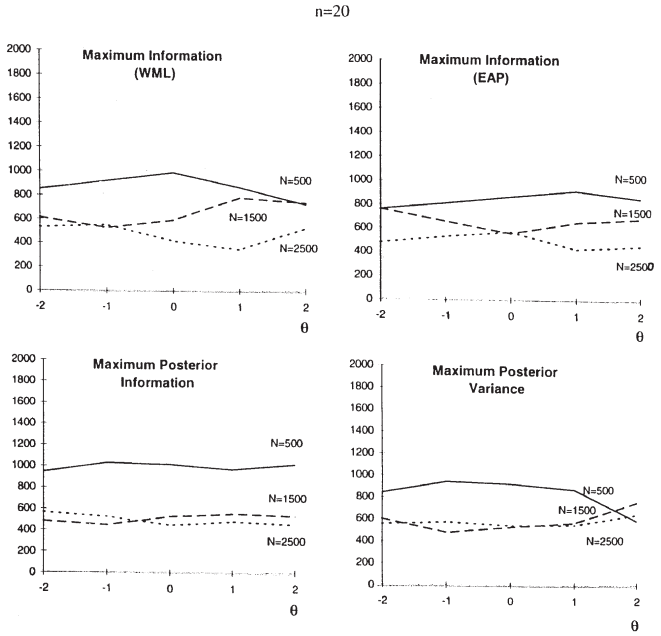

FIGURE 2. *Number of items in the adaptive tests selected from the sections in the pool calibrated on N = 500, 1,500, and 2,500 examinees for the various item-selection criteria (n = 20)*

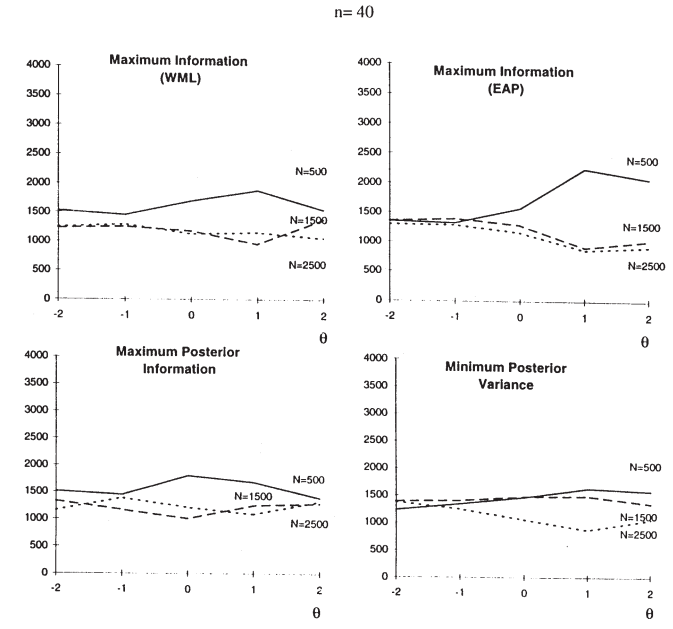

FIGURE 3. *Number of items in the adaptive tests selected from the sections in the pool calibrated on N = 500, 1,500, and 2,500 examinees for the various item-selection criteria (N = 40)*

Though not reported in these figures, the values of the discrimination parameters, a_i, for the items selected were broken down into sets of items with $a_i < .7$ and $a_i \geq .7$. This distinction corresponds to items with discrimination values below and above the average value for the items in the pool (see Table 1). However, for nearly all θ values and item-selection criteria, items with values for a_i in the lower category were never chosen. The only exception were a few cases with low θ values for the maximum-information criterion. These results remind us of a point well known in the practice of adaptive testing: Due to the presence of low discriminating items, the effective size of the item pool is generally much smaller than the number of items present in the pool.

In figures 4–6, each curve represents the mean absolute error in the ability estimates as a function of θ for the item pools calibrated on samples with sizes of N = 500, 1,500, and 2,500 examinees, the mixture of these sample sizes used above, and the true parameter values ($N = \infty$). For n = 10, the U-shaped curves typical of a short adaptive test with a prior for the ability parameter located at $\theta = 0$ were obtained. For n = 20 and 40, the curves became flatter, where the curves for the Bayesian item-selection criteria tended to be lower and flatter than those for the maximum-information criterion. Though the four criteria showed different degrees of capitalization on calibration error in Figures 1–3, the curves in Figures 4–6 were more homogeneous. the most conspicuous property of the latter, however, was a much larger variation in the mean absolute error between the different calibration samples at the higher part of the ability scale. at this part of the scale, the size of the mean absolute errors was inversely related to the size of the calibration sample. This result is due to the larger supply of difficult items in the pool (see Table 1). As a consequence, the item-selection ratio at this part of the scale is considerably smaller, and the tendency to capitalize on item parameter estimation errors is much stronger.

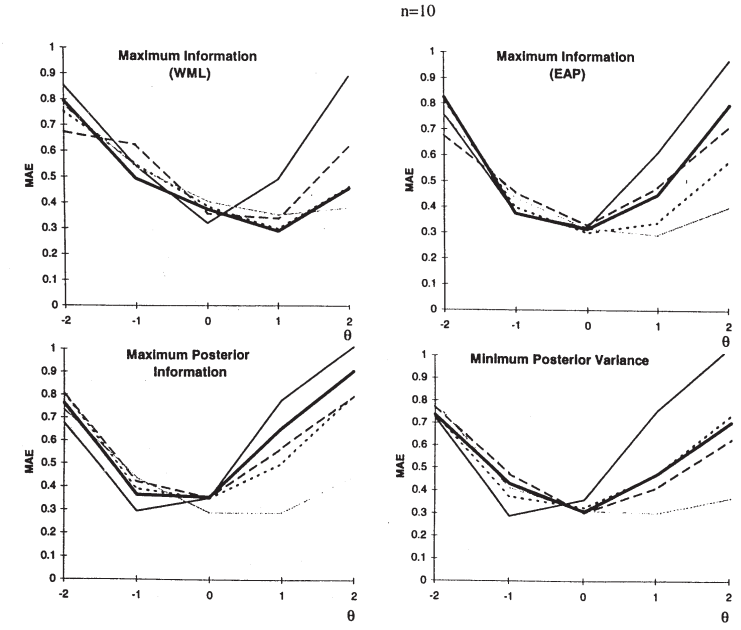

FIGURE 4. *Mean absolute error in the ability estimates for item pools calibrated on N = 500 (solid curve), 1,500 (dashed curve), 2,500 (dotted curve), a mixture of these sample sizes (bold curve), and N = ∞ examinees (grey curve) for the four item-selection criteria (n = 10)*

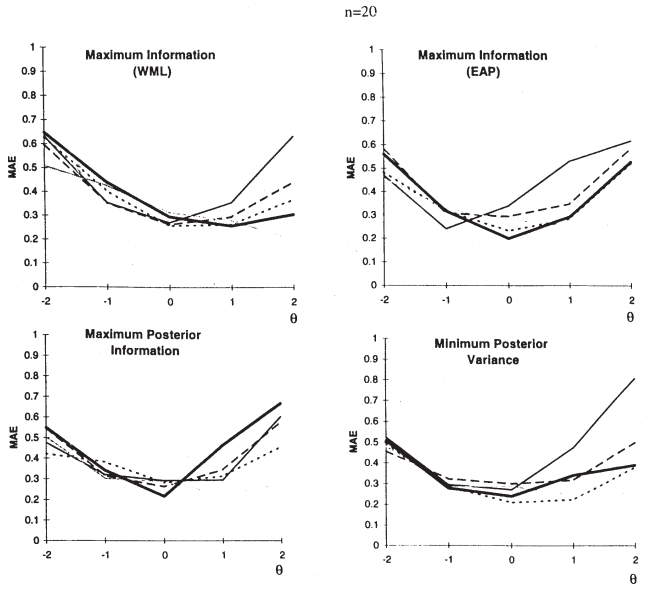

FIGURE 5. *Mean absolute error in the ability estimates for item pools calibrated on N = 500 (solid curve), 1,500 (dashed curve), 2,500 (dotted curve), a mixture of these sample sizes (bold curve), and N = ∞ examinees (grey curve) for the four item-selection criteria (n = 20)*

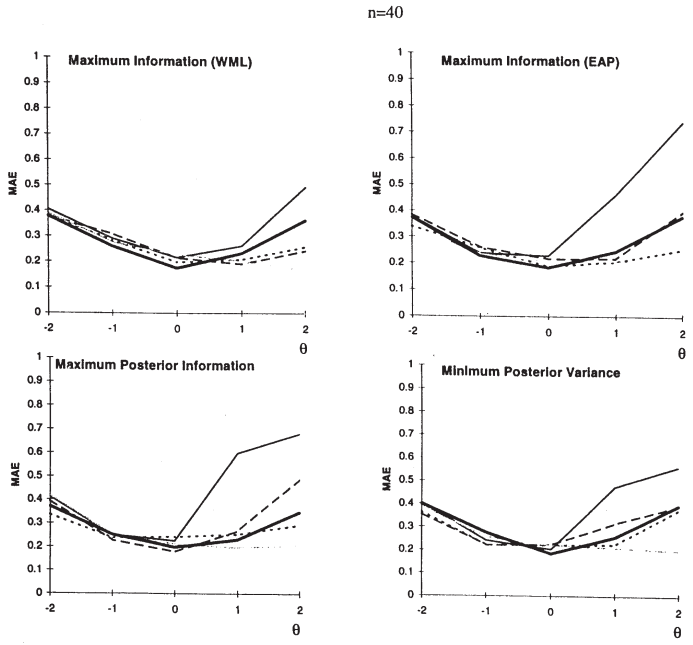

FIGURE 6. *Mean absolute error in the ability estimates for item pools calibrated on N = 500 (solid curve), 1,500 (dashed curve), 2,500 (dotted curve), a mixture of these sample sizes (bold curve), and N = ∞ examinees (grey curve) for the four item-selection criteria (n = 40)*

The effect of the item-selection ratio is also shown in Figure 7. For an item pool with size k = 40, that is, a large item-selection ratio, capitalization on calibration errors is not expected to occur and the mean absolute error in the ability estimates is high at the lower end of the scale and smaller at the higher end. The curves reflect the fact that the majority of the items were relatively difficult. If the size of the item pool increases, and the item-selection ratio decreases, the curves for the smaller calibration samples deteriorated at the higher end of the scale, whereas the curve for the true parameter values further improved. This increase in differences between the curves for the various sample sizes at the high end of the scale across the four panels in this figure is therefore expected to be due to capitalization on calibration error.

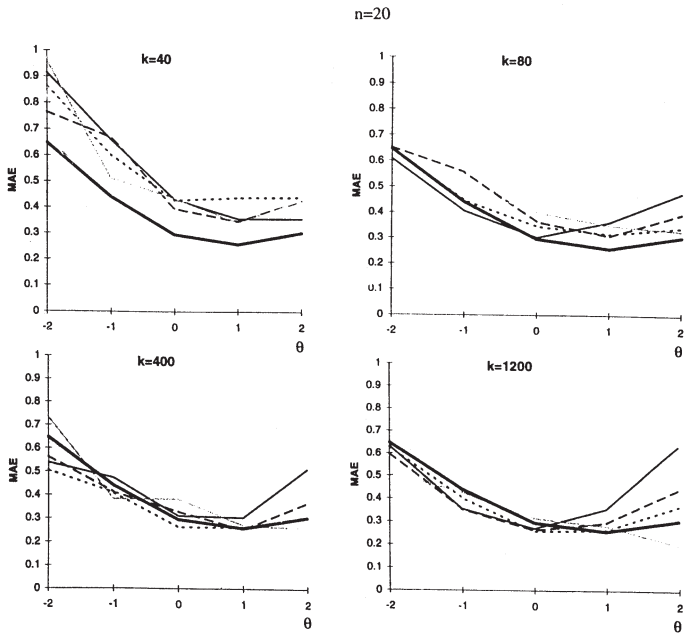

FIGURE 7. *Mean absolute error in the ability estimates for item pools calibrated on N = 500 (solid curve), 1,500 (dashed curve), 2,500 (dotted curve), a mixture of these sample sizes (bold curve), and N = ∞ examinees (grey curve) for pool sizes of k = 40, 80, 400, 1,200 items (maximum-information criterion with the weighted maximum likelihood estimation of ability (n = 20)*

Conclusion

The general picture emerging from this example is that capitalization on calibration error does occur in adaptive testing and that its most important determinant is the item-selection ratio. Item pools and test lengths of various sizes were used to study the effects of this ratio on the ability estimates. However, because the item pools were generated from an empirical data set, difficult items were overrepresented, the result being an actual item-selection ratio smaller than expected at the higher end of the ability scale.

This unexpected result showed that the composition of the item pool is an important factor interacting with the effect of capitalization on errors in the item parameters on the errors in the ability estimates. Large numbers of items for certain θ values–intuitively an attractive feature of an item pool—is *not* a desideratum if the calibration sample is small.

References

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

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. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

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

Lord, F. M. (1983). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York: Academic Press.

Lord, F. M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351–356.

Tang, K. L., Way, W. D., & Carey, P. A. *The effect of small calibration sample sizes on TOEFL IRT-based equating* (TOEFL Technical Report TR-7). Princeton, NJ: Educational Testing Service.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty on item parameter estimation on ability estimates. *Psychometrika, 55,* 371–390.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 210-216.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22,* 259–270.

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

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: primer.* Hillsdale, NJ: Erlbaum.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.