Graduate Management Admission Council®

An Automatic Online Calibration Design in Adaptive Testing

Guido Makransky

University of Twente

Presented at the Real-Time Analysis Paper Session, June 2, 2009

Abstract

An accurately calibrated item bank is essential for a valid computerized adaptive test. However, in some settings, such as occupational testing, there is limited access to examinees for calibration. As a result of the limited access to possible examinees, collecting data to accurately calibrate an item bank in an occupational setting is usually difficult. In such a setting, the item bank can be calibrated online in an operational setting. This study explored three possible automatic online calibration strategies, with the intent of calibrating items accurately while estimating ability precisely and fairly. That is, the item bank is calibrated in a situation where examinees are processed and the scores they obtain have consequences. A simulation study was used to identify the optimal calibration strategy. The outcome measure was the mean absolute error of the ability estimates of the examinees participating in the calibration phase. Manipulated variables were the calibration strategy, the size of the calibration sample, the size of the item bank, and the item response model.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Makransky, G. (2009). An automatic online calibration design in adaptive testing. 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

Guido Makransky, Ingerslevsgade 132. 3 th. Copenhagen 1705, Denmark.
Email: guidomakransky@gmail.com

An Automatic Online Calibration Design in Adaptive Testing

The past 15 years have seen a steady increase in the use of online testing applications in a variety of testing settings. Computers can be used to increased statistical accuracy of test scores using computerized adaptive testing (CAT; Van der Linden & Glas, 2000a). The implementation of CAT is attractive because research indicates that CATs can yield ability estimates that are more precise (Rudner, 1998, Van der Linden & Glas, 2000a), can be more motivating (Daville, 1993), easier to improve (Linacre, 2000, Wainer, 2000), and take a shorter period of time to complete (Rudner, 1998; Wainer, 2000) than traditional tests. Although CATs have been widely implemented within large scale educational testing programs, the use of CATs in other settings such as in occupational testing has been limited because of several practical challenges.

One of the major obstacles to cost-effective implementation of CAT is the amount of resources needed for item calibration. Large testing programs have been able to overcome this problem with the availability of extensive resources. Nevertheless, there has been broad interest in investigating procedures for optimizing the calibration process (e.g., Berger 1991; 1992; 1994; Berger, King & Wong, 2000; Jones & Nediak, 2000); Lima Passos, Berger; 2004; Stocking 1990). Unfortunately, this research is based on the assumption that a large number of examinees is available in the development phase of a test. However, this is not the case in many applied settings. In reality, the lack of available examinees is one of the greatest challenges in the development phases of a test in an occupational setting. This is the case because the organizations that purchase an occupational test are usually unwilling to invest time and resources in letting their employees take a test unless they can use the results. To circumvent this problem, test developers usually access examinees from a context other than the one in which the test is to be used, that is, they access a low-stakes sample. The use of a low-stakes calibration sample comes with several limitations. First, there is evidence that large motivational differences exist between examinees in low-stakes calibration samples and the intended population of examinees (Wise & DeMars, 2006). These motivational differences introduce bias in the estimation of item parameters in the calibration phase, which will result in biased test scores. Further, the use of a separate sample usually means extra resources in terms of time and money in test development.

The resources required for item calibration would be reduced if a test could be calibrated and implemented for the intended population as quickly and fairly as possible. This would make it attractive for possible customers to be involved in the calibration process because they could use the results. Therefore, it is worthwhile to identify designs that make it possible to simultaneously calibrate items and estimate ability, while treating examinees fairly. The present study differs from previous studies in that this is an investigation of the problem of calibrating a set of items where there is no previously available information, with the practical constraint of maintaining fairness in test scoring. The purposes of this paper are to discuss calibration strategies that will make it more practical and cost effective to develop and implement CATs in occupational settings, and to report on a simulation study conducted to choose an optimal strategy.

The Model

The present study was carried out in the framework of item response theory (IRT). The fundamental concept of IRT is that each test item is characterized by one or more parameters and

each examinee is characterized by a single ability parameter. The probability that a given examinee answers a given item correctly is given by a function of both the item's and the examinee's parameters. Conditional on those parameters, the response on one item is independent of the responses to other items. The IRT model used in this study, is the two-parameter logistic, or 2PL model,

$$P_i(\theta) = \frac{1}{1 + \exp(-a_i(\theta - b_i))} \tag{1}$$

(Birnbaum, 1968). Here $P_i(\theta)$ is the probability of a correct response for item i, θ is the examinee's ability, and a_i and b_i are item parameters. a_i is called the discrimination and b_i the difficulty parameter. A specific form of this model is the one-parameter logistic or1PL model. In the 1PL model the assumption is made that all items have the same discrimination parameter.

Calibration pertains to the estimation of the item parameters a_i and b_i from response data, say data from a calibration sample. In the operational phase of CAT, the item parameters are considered to be known and the focus becomes the estimation of θ. In IRT, θ can be estimated using several different strategies. The weighted maximum likelihood estimator derived in Warm (1989) was used to estimate θ in this study. This method is attractive because of its negligible bias (Van der Linden & Glas, 2000b).

What differentiates CAT from traditional tests is that items are selected optimally by an item selection algorithm that finds the next available item from the item bank that provides the most information about the examinee. A selection function that is often used in item selection for CAT is Fisher's information function. For dichotomously scored items, the information function has the following form:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{2}$$

where $P_i(\theta)$ is the response function for item i, $P'(\theta)$ its first derivative with respect to θ, and $Q_i(\theta) = 1 - P_i(\theta)$. In CAT, the item is selected that has the maximum information in the item pool at $\theta = \theta^*$, where θ^* is the current θ estimate for the examinee (Van der Linden & Glas, 2000b). Maximization of information minimizes the estimation error of θ.

Calibration Strategies

This study investigated the online calibration of an item bank when there is no available information about item parameters at the beginning of the testing process. Therefore, the most equitable way to select items during the initial phase of testing is to administer items randomly. Although random item administration does not guarantee tests with equal difficulty levels, it does ensure that there are no systematic differences in difficulty which would result in unfairness. Then, once sufficient data become available, optimal item selection can be carried out with Fisher's information function. The objective of this paper is to investigate how to progress from random to optimal item selection in a fair and effective manner. The next section describes three plausible calibration strategies for attaining this goal.

Two-Phase Strategy

In this strategy, labeled P2, items are administered randomly up to a given number of examinees. For the remaining examinees the items are calibrated and administered optimally in the form of a CAT. In the random phase, tests are scored with the assumption that all items have a difficulty parameter equal to 0 (that is, $b_i = 0$), and in the optimal phase tests are scored based on the item parameters obtained in the random phase. The reason for the scoring rule in the random phase is to obtain scores that are on the same scale as in the optimal phase. This scoring rule is analogous to the scoring rule used in classical test theory, where a proportion-correct score is computed assuming that all items have the same weight. Here this score is simply converted to a score on the θ scale. The clear transition from one phase to the next means that stakeholders can be informed about the current precision of the test, and policy decisions about how the test should be used can be clearly defined based on the level of precision. The transition is made when the average number of item administrations is above some predefined value T. The optimal transition point T from the random to the optimal phase was one of the topics in this study.

Multi-Phase Strategy

An alternative strategy labeled M consists of more than two phases. As in the previous strategy, the items are calibrated at the end of each phase. Table 1 illustrates an example with the five phases that the design follows. In Phase 0, all item selection is random and θ is estimated with the assumption that $b_i = 0$. As in the previous strategy, also here the transition is made when the average number of item administrations is above some predefined value T. In the next phase, labeled Phase 1, the first three parts of the test are random, and the final part is CAT using the item parameter estimates from data collected in the previous phase. A transition takes place when the average number of administrations over items has doubled. In general, a transition takes place when this average exceeds (Phase + 1) × T. This continues until the final phase, where all of the items are administered optimally and the item bank is calibrated.

Table 1. Phases of the Multi-Phase Strategy

Phase	Part 1	Part 2	Part 3	Part 4
0	Random	Random	Random	Random
1	Random	Random	Random	CAT
2	Random	Random	CAT	CAT
3	Random	CAT	CAT	CAT
4	CAT	CAT	CAT	CAT

The motivation for the strategy is as follows: In phase 0, the amount of uncertainty regarding the item parameters and the person parameters is too high to allow for optimal item selection. In fact, this high uncertainty might introduce bias because the uncertainty estimate in item parameters and θ could compound the error in the θ estimate. Therefore, items are administered randomly. After the random part, θ is estimated using the item parameters obtained in the previous phase, and this estimate serves as an initial estimate for the adaptive part. In later phases, it is assumed that the parameters are estimated with sufficient precision to support optimal item selection. The inclusion of an adaptive part at the end makes the test more effective

in terms of scoring ability and in terms of calibrating items. As with the P2 strategy there is a clear transition point between phases in this strategy.

Continuous Updating Strategy (C)

Labeled C, this strategy is analogous to the previous two strategies in that items are administered randomly and tests are scored with the assumption that $b_i = 0$ in the first phase. An item becomes eligible for CAT if the number of administrations of the item is above a transition point labeled T. The proportion of CAT in a test is proportional to the number of eligible items in the item bank. In the final phase where all item selection is optimal, items are calibrated after each exposure and tests are scored based on the parameters computed after the latest administration of the items. Therefore, the precision of the θ estimates is continuously improved.

The three calibration strategies represent a sample of possible designs on a continuum ranging from one extreme where items are calibrated at a single point in time, to the other extreme where items are calibrated constantly after each exposure, once the items become eligible for CAT.

Simulation Studies

To investigate which of the considered calibration strategies leads to the lowest overall mean absolute error (MAE) in the estimation of θ, simulation studies were conducted. The studies were designed to measure the impact of each of the three strategies across a variety of conditions by varying the following variables:

1. The transition point T from one phase to the next. These points were varied as $T = 10, 25, 50, 100, 200$ item administrations.
2. The calibration sample sizes, which were varied as $N = 250, 500, 1{,}000, 2{,}000, 3{,}000, 4{,}000$.
3. The IRT model, varied as the 1PL model and the 2PL model.
4. The size of the item bank, varied as $K = 100, 200, 400$ items.

Upper and lower baselines were also simulated to compare the precision of the simulation strategies to external criteria. MAE for an optimal test administered with a completely calibrated item bank, labeled O, was set as a lower baseline. This was simulated by calibrating items using strategy P2 with a transition point of 4,000. The precision of a test administered randomly with all items having difficulty parameters of 0.0 was set as an upper baseline. This procedure is labeled R.

Method

The focus of this study was to assess how accurately θ is estimated while in the calibrating phase of the test. Once the number of examinees becomes large and the item bank is accurately calibrated, it is expected that different calibration designs result in similar precision, so then the calibration design is no longer of interest. Therefore, it was important to differentiate the calibration sample from the post-calibration sample of examinees. A calibration sample of 4,000 examinees was set in this study.

The examinees' θ parameters were drawn from a standard normal distribution. An item bank was simulated by drawing item difficulty parameters from a standard normal distribution, and item discrimination parameters from a lognormal distribution with an expectation of 1. After each phase, items were calibrated under either the 1PL or 2PL model using the method of

marginal maximum likelihood estimation (Bock & Aitkin, 1981). Optimal item selection was implemented using maximal expected information. The item parameters were the current estimates at that point in the design of the strategy, and the test length was set at 20**.** MAE was computed as the mean absolute difference between the true θ drawn from the N(0,1) distribution and the θ estimated by the weighted maximum likelihood procedure. The MAE for each strategy was then calculated by averaging across all examinees to give an estimate of the global precision of the strategy.

In addition to global precision, it was also of interest to investigate the precision with which a certain examinee's score was estimated. This so-called local precision was measured at specific points on the θ continuum ($\theta = -2, -1, 0, 1, 2$), to give an estimate of the precision with which an examinee with a specific θ could be expected to be assessed within each condition. Therefore, after each phase 4,000 examinees were simulated at each of the five θ values, and the MAE was computed for each of the five θ values.

Results

Global precision and optimal transition points. The first research question investigated was the optimal point at which item selection should transition from one phase to the next in each of the three calibration strategies. Five conditions were investigated ($T = 10, 25, 50, 100, 200$) for the 1PL and 2PL models. The results are shown in Table 2.

The table gives the MAE obtained for the three calibration strategies as well as a completely random (R) and completely calibrated test (O), for a calibration sample size of 4,000, with item bank sizes of 100, 200 and 400 ($K = 100, 200, 400$), using the 1PL and 2PL models. A comparison of the MAE for the three strategies indicated that the C strategy consistently resulted in the best θ estimates across all conditions.

The results for the 1PL model were consistent across the item bank sizes, and indicated that a transition point of 100 ($T = 100$) had the lowest MAE for the P2 strategy, $T = 50$ for the M strategy, and $T = 25$ for the C strategy. Therefore, the most effective transition point became lower as the number of calibration points for the strategy increased (from P2 to C). Note that for T = 10, the MAE of the P2 and M strategies was often above the MAE of the upper baseline (strategy R). This occurred because, in that case, the item parameters were calculated based on 10 observations only. Therefore these estimates of the item parameters were very poor and performed worse than the baseline estimate of $b_i = 0$.

The results for the 2PL model were similar to those for the 1PL, but they were not as consistent. Specifically a faster transition seemed to be optimal for the M strategy with larger item bank sizes. This finding seems to be a consequence of the M strategy taking a long time to transition through the five phases in the design with large item banks.

The general pattern in these findings is consistent with the hypothesis that a balance between efficiency and accuracy in terms of switching from one phase to the next is important. A quick transition resulted in a premature progression through the phases in each strategy, because item parameter estimates still had much error. Therefore, the use of an optimal item selection

Table 2. Comparison of the MAE for Different Transition Points Within Each Calibration Strategy

			MAE					
Model	Item Bank	Strategy		$T = 10$	$T = 25$	$T = 50$	$T = 100$	$T = 200$
1PL	$K = 100$	R	0.418					
		P2		0.489	0.420	0.404	**0.392**	0.394
		M		0.453	0.395	**0.389**	0.396	0.402
		C		0.381	**0.379**	0.380	0.381	0.392
		O	0.376					
	$K = 200$	R	0.418					
		P2		0.577	0.435	0.409	**0.393**	0.396
		M		0.417	0.397	**0.392**	0.408	0.420
		C		0.390	**0.382**	0.393	0.398	0.404
		O	0.381					
	$K = 400$	R	0.418					
		P2		0.533	0.439	0.406	**0.401**	0.414
		M		0.430	0.410	**0.405**	0.414	0.420
		C		0.400	**0.396**	0.397	0.397	0.413
		O	0.384					
2PL	$K = 100$	R	0.405					
		P2		0.475	0.388	0.353	**0.352**	0.361
		M		0.362	0.366	**0.358**	0.368	0.380
		C		**0.342**	0.345	0.353	0.355	0.366
		O	0.342					
	$K = 200$	R	0.405					
		P2		0.460	0.352	0.351	**0.349**	0.373
		M		0.366	**0.340**	0.346	0.381	0.394
		C		0.335	**0.324**	0.329	0.352	0.369
		O	0.323					
	$K = 400$	R	0.405					
		P2		0.450	0.368	**0.356**	0.362	0.416
		M		**0.344**	0.354	0.375	0.401	0.406
		C		0.339	**0.330**	0.348	0.366	0.414
		O	0.311					

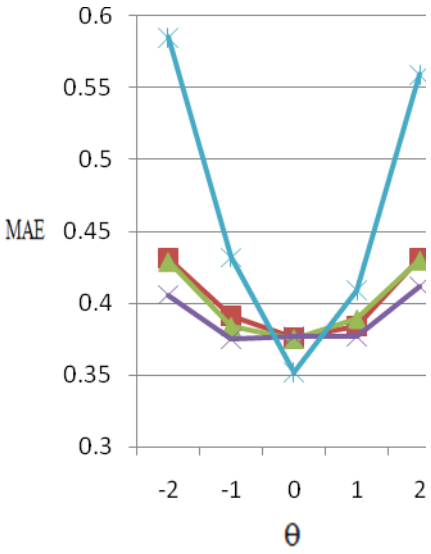
Note. Best results for each strategy within each condition are in boldface.

algorithm to administer items, assuming that the item parameters were accurate, resulted in inaccurate θ estimates. On the other hand, the slower progression through the phases resulted in loss of efficiency because the calibration procedure did not react quickly enough in switching to the next phase, even though item parameter estimates had stabilized. Since the results were similar across the different item bank sizes, and between the two models, transition points of $T = 100$, $T = 50$, $T = 25$ were used respectively, for the P2, M, and C calibration strategies in subsequent analyses for both the 1PL and 2PL models, in order to have comparable results across settings.

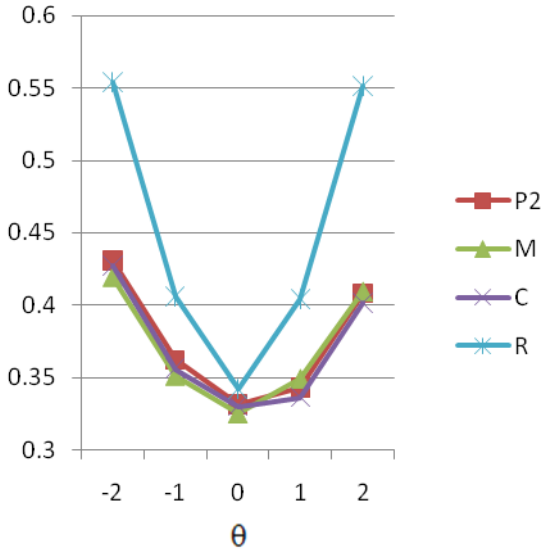
Local comparison of the calibration strategies. In addition to global precision, the local precision of the three strategies for specific points on the θ scale was investigated. A comparison of these and random item administration with $b_i = 0$ (R) as a baseline is presented in Figure 1.

Figure 1. MAE at Specific Points on the θ Continuum

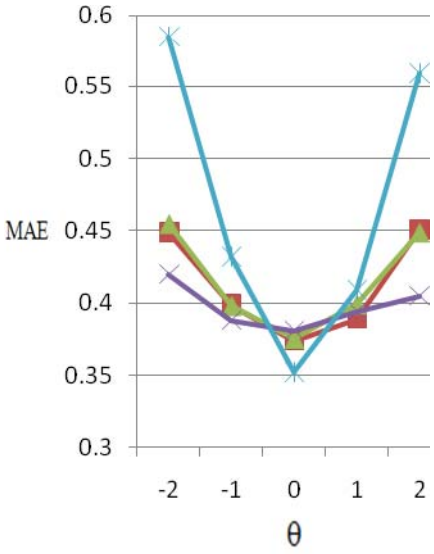
a. Strategies P2, M, C and R, 1PL Model, K = 100



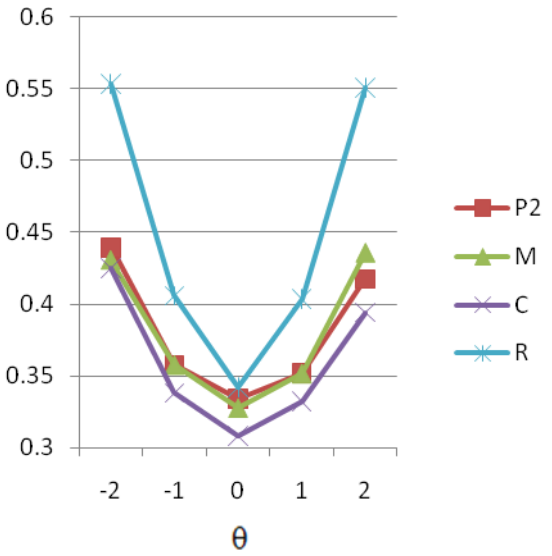
b. Strategies P2, M, C and R, 2PL Model, K = 100



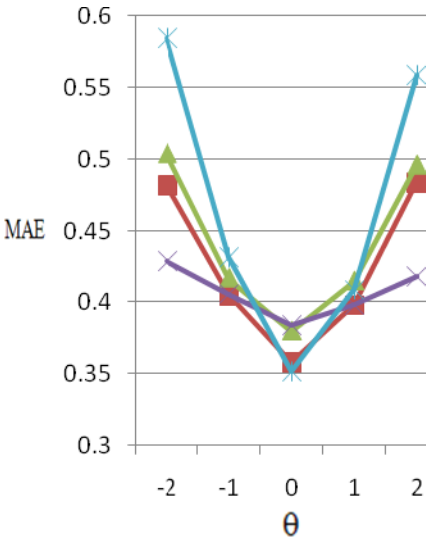
c. Strategies P2, M, C and R, 1PL Model, K = 200



d. Strategies P2, M, C and R, 2PL Model, K = 200



e. Strategies P2, M, C and R, 1PL Model, *K* = 400



f. Strategies P2, M, C and R, 2PL Model, *K* = 400

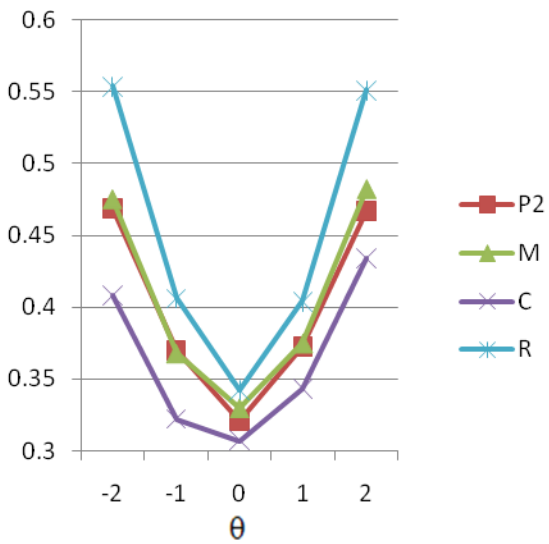
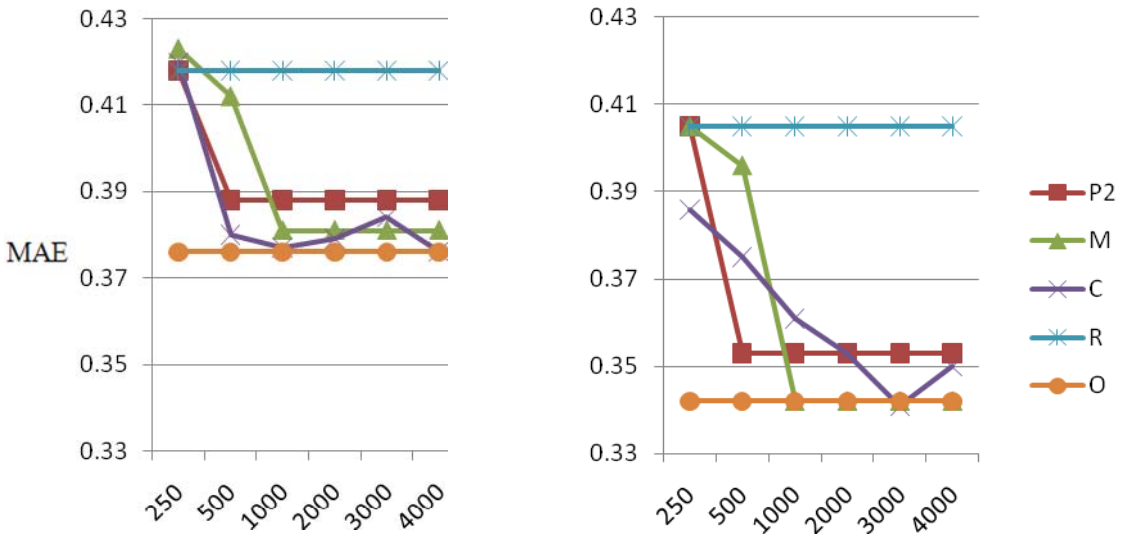


Figure 1 illustrates the local precision of the three strategies with the 1PL model on the left side and the 2PL model on the right side, for item bank sizes of 100, 200, and 400 items. The horizontal axis represents the ability level of the examinees at five points on the θ scale (–2, –1, 0, 1, 2), and the vertical axis represents the MAE across the first 4,000 examinees within each design. For the 1PL model, the graph shows that the C strategy measured θ more precisely than the other strategies at extreme θs, while all three strategies performed fairly equally at $\theta = 0$. The use of random item administration with item parameter estimates of $b_i = 0$ performed well at $\theta = 0$; however, this method performed much poorer at extreme θ levels. For the 2PL model, the three strategies performed quite similarly with a smaller item bank, but the C strategy performed better than the other two as the item bank size became larger. All three strategies also performed better than random item administration for the 2PL model, with the largest differences occurring at extreme θ values.

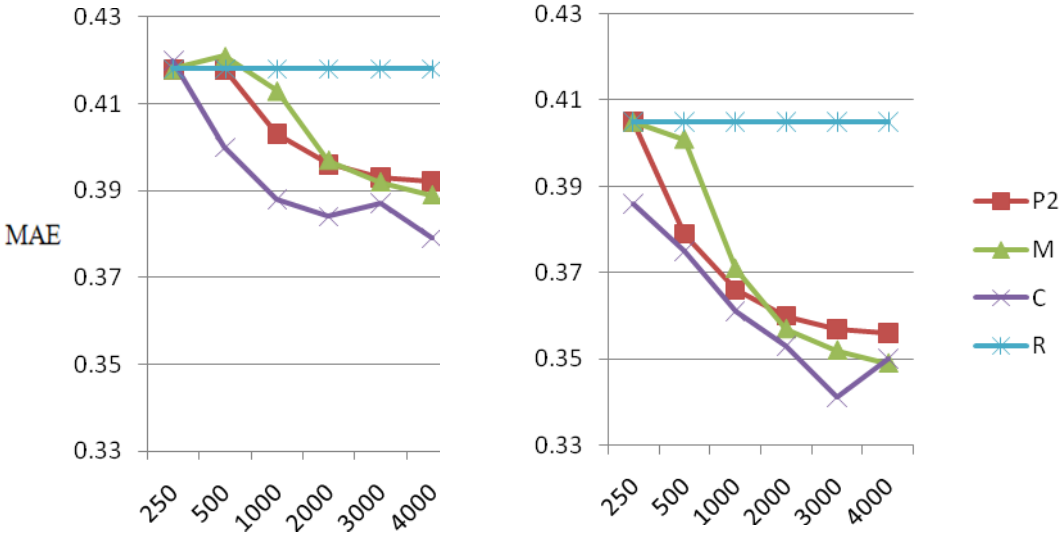
A comparison of the strategies at different points in the calibration process. The next research question investigated was the precision of the strategies for settings with a limited number of examinees. In this section the examples are limited to an item bank size of 100, because the general results across the different item bank sizes led to similar conclusions. Figures 2a and 2b display the specific precision for each strategy at a particular point in the calibration process. In other words, these figures present the results for how accurately the particular test estimates θ for the n^{th} examinee in the calibration design. This provides information about the point at which a test can be confidently used in a high-stakes situation. The horizontal axis represents the n^{th} examinee in the calibration design, and the vertical axis shows the MAE for the three strategies, as well as random item administration (R), and a fully calibrated test (O).

Figure 2. Precision of R and O Strategies Examinees Number 250, 500, 1,000, 2,000, 3,000 and 4,000 for the 1PL Model, and the 2PL

a. Specific Precision, 1PL Model **b. Specific Precision, 2PL Model**



c. Cumulative Precision, 1PL Model **d. Cumulative Precision, 2PL Model**



The results indicate that strategy C performed nearly as well as a fully calibrated test after as few as 500 examinees for the 1PL model; it took strategy M 1,000 examinees to reach a similar level of precision. Strategy P2 never reached the same precision as a fully calibrated test, which implies that the P2 strategy needs to be supplemented with additional calibration points later in the design in order to reach the same level of accuracy. The results for the 2PL model were similar to the 1PL model, with the exception that the C strategy took a longer time to reach precision estimates comparable to a completely calibrated test.

These results consider the accuracy of a given examinee at a particular point in the

calibration process. The Figures 2c and 2d present the cumulative precision of each strategy, which is the average precision with which an examinee is assessed in the calibration phase of the test, for different size calibration samples. The figure plots the average MAE of the sample on the vertical axis, based on the number of examinees in the calibration sample on the horizontal axis. The results were similar for the 1PL and 2PL models, in that the C strategy performed considerably better than the other two strategies and random item administration. The difference was evident after the number of examinees in the calibration sample reached 500 for the 1PL model, and after as few as 250 for the 2PL model. The M and P2 strategies resulted in θ estimates that were considerably better than random item administration; however. the calibration sample had to be at least 1,000 before a significant difference was evident. The difference between the precision of the three strategies decreased as the calibration sample became larger, suggesting that the benefits of using the C strategy are highest when there is a limited number of examinees.

Item exposure. The calibration strategies have been compared in terms of how accurately θ is assessed in the calibration process. However, another purpose of this study was to identify a calibration method that would calibrate the entire item bank. Therefore, it was important to investigate the frequency with which items were administered using the calibration strategies in the two models. Table 3 displays the number of times items were administered in the three calibration strategies, for item bank sizes of $K = 100$ and $K = 400$, in a calibration sample of 4,000 examinees. The results for the 1PL model are presented in the upper portion, and the 2PL model in the lower portion of the table.

Table 3. Number of Times Items were Administered for Each Strategy Within Each Model

			Number of Administrations						
Model	Item Bank	Strategy	< 100	100-199	200-399	400-599	600-799	800-999	> 1000
1PL	K = 100	P2	0	2	13	21	31	18	15
		M	0	0	6	28	31	23	12
		C	0	0	0	32	39	6	23
	K = 400	P2	0	286	103	6	1	0	4
		M	0	316	64	8	0	0	12
		C	0	262	142	0	0	0	0
2PL	K = 100	P2	12	30	11	8	4	5	30
		M	0	45	7	8	5	6	29
		C	39	6	6	10	6	5	28
	K = 400	P2	136	198	22	11	15	6	12
		M	1	355	19	7	8	8	10
		C	320	17	18	8	6	4	27

Table 3shows a fairly uniform administration of items for all three calibration strategies for the 1PL model. Item administration for the 2PL model was highly uneven for the P2 and C strategies, but fairly balanced for the M strategy. In the C strategy, 39%, and 80% of the items were administered fewer than 100 times, for item banks consisting of 100 and 400 items respectively.

Discussion

The purpose of this study was to investigate which of three possible calibration strategies would result in the lowest MAE in θ estimation, and lead to a uniform administration of items in a setting where examinees' ability could be assessed throughout the calibration design. The benefits of the three designs were tested in terms of several possible conditions. In general, all strategies performed well across the different conditions and seem to be viable options when calibrating an item bank effectively and fairly, so as to use the test in a high-stakes setting as quickly as possible.

The C strategy consistently outperformed the other two strategies across all conditions. In fact, θ was estimated nearly as well as in a fully calibrated test after as few as 500 examinees in a test consisting of 20 items and an item bank consisting of 100 items for the 1PL model. A weakness of this strategy was the non-uniform administration of items with the 2PL model, which lead to the calibration of a few items at the expense of others. The M strategy might be preferred in settings where the 2PL is used, because this strategy resulted in a more uniform administration of items with both models. However, a larger number of examinees were required before the precision in θ estimation increased, which made this strategy ineffective with large item bank sizes. The P2 strategy generally resulted in a lower level of precision compared to the other two, because items were calibrated only at one point. An alternative method would be to use the P2 strategy with follow-up calibrations instead of simply calibrating one time. The use of random item selection with $b_i = 0$ for all parameters at the beginning of each strategy, led to good θ estimates for examinees with θ estimates near the mean; however, this method was inaccurate at estimating examinees with extreme θ values due to a consistent shrinkage toward the mean.

In a context where stakeholders need to know the level of precision in the test in order to make procedural decisions about how the test should be used, it might be important that examinees within the same phase are given the same probability of success. Here the P2 or the M strategy would be preferred over the C strategy because the precision in the C strategy is continuously improved.

The C and P2 strategies resulted in a non-uniform administration of items for the 2PL model, because the item selection algorithm in the 2PL model quickly resorted to selecting the items with high discrimination parameters at the expense of the other items. This resulted because the discrimination parameter has a multiplicative effect on the information for the items for the 2PL model, which leads to the selection of items with greater information at specific points on the θ scale, over items that provide information across a broader area. This can be efficient when there is little error in θ and item parameter estimates; however, it is not optimal at the beginning of a test when there is a lot of insecurity concerning an examinee's θ, and is undesirable when there is error in the item parameters. The use of the 2PL model for these strategies could be a disadvantage because items can receive a small discrimination parameter by chance due to inconsistent answering in a small test taker population. Therefore, good items might never get the opportunity to be accurately calibrated and used in the test with the 2PL model, which would result in a waste of resources for the test development organization. The optimal selection of items in the development phases of a test with the 2PL model could also be an advantage, however, in settings where there is an abundant number of items and it does not matter if some

items are never used, because the algorithm in the 2PL model concentrates on calibrating the items that are likely to be the best and most frequently used in the test.

A study by Van der Linden and Glas (2000b) found dramatic impact of capitalization on estimation errors on θ estimation using the 2PL model with a fully calibrated test. They highlighted four solutions for controlling the capitalization of error in θ estimation: cross validation, controlling the composition of the item pool, imposing constraints, and using the 1PL model. The final two are possibilities for the current context. Imposing an exposure constraint would lead to a more uniform administration of items; however, the constraints would also limit the efficiency of the item selection algorithm**.** In the context of the 1PL model all three calibration strategies resulted in improved θ estimates, in addition to a uniform calibration of items. The results suggest that the 1PL model could be used in selecting items for the calibration phase of the test, and then once items have been accurately calibrated, the selection algorithm could switch to the 2PL model.

The results of the study provide viable calibration design options for test development organizations that find it difficult to get test takers in the development phases of a test. In these settings, these calibration strategies offer more cost effective and practical methods for developing large item banks, which makes it more attractive for smaller test development organizations to take advantage of the benefits of CAT. All three methods have the advantage over traditional booklet calibration designs that they offer the possibility to assess test takers' ability throughout the calibration of the test. This makes it more attractive for organizations that purchase occupational tests to become involved in the development phases of the test because the results can be used.

Future research could investigate the consequences of using the 2PL model with item exposure constraints to investigate if it can lead to a uniform calibration of items while simultaneously estimating ability accurately. In this study, the assumption was made that items fit the model that was used; future research could also estimate the consequences of bad items by varying the degree to which the items fit the model. Finally, methods for filtering and assessing fit in items during the calibration process could be considered.

References

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, *15*, 283-306.

Berger, M. P. F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*, *57*, 521-538.

Berger, M. P. F. (1994). D-optimal designs for item response theory models. *Journal of Educational Statistics*, *19*, 43-56.

Berger, M. P. F., King, C. Y. J., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika, 65*, 377-390.

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

Bock R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.

Daville, C. (1993). Flow as a testing ideal. *Rasch Measurement Transactions, 7(3)*.

Jones, D. H., & Nediak, M. S. (2000). *A simulation study of optimal on-line calibration of testlets using real data* (RUTCOR Research Report). New Brunswick, NJ: Rutgers University, Faculty of Management and RUTCOR.

Lima Passos, V., & Berger, M. P. F. (2004). Maximin calibration designs for the nominal response model: An empirical evaluation. *Applied Psychological Measurement*, *Vol. 28*, 72-87.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come.* MESA Memorandum. No. 69.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Pædagogiske Institut.

Rudner, L. (1998). *An applied study on computerized adaptive testing*. Rockland, MA: Swets & Zeitlinger.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational & Behavioral Statistics, 21*, 405-414.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2000a). *Computerized adaptive testing. Theory and practice.* Dordrecht: Kluwer Academic Publishers.

Van der Linden, W. J., & Glas, C. A. W. (2000b). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35-53.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, *55*, 461-475.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer.* Second edition. Hilsdale, NJ: Lawrence Erlbaum Associates.

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

Wise, S. L., & DeMars, C. E. (2006). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.