An Automatic Online Calibration Design in Adaptive Testing[1]

Guido Makransky[2]

Master Management International A/S and University of Twente

Cees. A. W. Glas

University of Twente

[1] Conflict of Interest Disclosure: There is no immediate conflict of interest related to this paper. The first author's doctoral project is sponsored by a grant from the Danish government and Master Management International A/S.

[2] Correspondence regarding this manuscript should be addressed to Guido Makransky, Gydevang 39-41, Alleroed 3450, Denmark, email: guidomakransky@gmail.com.

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 test takers for calibration. As a result of the limited access to possible test takers, 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 test takers 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 test takers 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.

Key Words: *computerized adaptive testing, item bank, item response theory, online calibration*

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 increase 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, because of the large item banks needed in CAT. Large testing programs have been able to overcome this 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 test takers 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 test takers is one of the greatest challenges in the development phases of a test in an occupational setting. This is the case because the companies 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 test takers 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 test takers in low stakes calibration samples and the intended population of test takers (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 test takers 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. This problem is common for test development companies that are interested in developing a new CAT when there is no previously available paper version of the instrument.

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 small testing programs, and to report on a simulation study that was conducted to choose an optimal strategy. More specifically, the paper investigates three different calibration strategies for calibrating an item bank from scratch, with the primary objectives of calibrating items in a fair and effective manner, while providing accurate ability estimates throughout the calibration design.

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

[illegible]

(Birnbaum, 1968). Here *P$_i$(_)* is the probability of a correct response for item *i*, _ is the test taker's ability, and a_i and b_i are item parameters. Further, a_i is called the discrimination and b_i the difficulty parameter. A specific form of this model that is also used in this study is the one-parameter logistic or1-PL model. In the 1-PL model the assumption is made that all items have the same discrimination parameter. The 1-PL and 2-PL models are viable alternatives to the 3-PL model because the guessing parameter in the 3-PL model can be difficult to estimate in small sample sizes, as those used in this study. An additional reason for not using the 3-PL model in this study is that a CAT algorithm is used to administer items from an early stage in the calibration strategies described in this study. Therefore, the chances of guessing are not as high because the ability of the respondent is matched with the difficulty of the item.

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 _. In IRT _ can be estimated using several different strategies. The weighted maximum likelihood estimator derived in Warm (1989) was used to estimate ability 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 test taker. A selection function that is often used in item selection for CAT is Fisher's information function. For an introduction regarding Fisher information and alternative criteria for item selection, refer to Wainer (2000) or van der Linden and Glas (2009a). For dichotomously scored items, the information function has the following form:

[illegible]

$$\overline{\qquad\qquad}\,,$$

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

Calibration Strategies

This study investigated the online calibration of an item bank where there was 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 was 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. A viable calibration strategy would be able to progress from random to optimal item selection in a fair and effective manner. The next section describes three plausible calibration strategies that were evaluated in this study.

Two-Phase Strategy

In this strategy, labeled P2, items are administered randomly up to a given number of test takers. For the remaining test takers 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 _ 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 ability 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.

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 ability could compound the error in the ability estimate. Therefore, items are administered randomly. After the random part, ability 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.

-- insert Table 1 here--

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 optimally administered items in a test is proportional to the number of eligible items in the item bank. Therefore, during this phase the first part of the test is random, and the final part is CAT where items are calibrated after each exposure and tests are scored based on the parameters computed after the latest administration of the items. In the final phase all item selection is optimal and the items are calibrated after each exposure, therefore, the precision of the _ estimates is continuously improved.

The three calibration strategies were chosen because they 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. The P2 strategy is similar to a typical random item administration calibration design where the items are calibrated at a single point, with the difference that ability scores are reported up to that point. Therefore, this strategy is the easiest to implement and can be considered a control strategy for comparison purposes. The P2 and M strategies have the advantage that test takers within the same phase are given the same probability of success. This can help define policy decisions about how the test should be used, based on the level of precision in the test. The C strategy has the advantage that changes in the calibration sample can be quickly detected because calibration occurs continuously. This would make it easier for test developers to detect mistakes in the items, and would make it possible to get a rough measure of the characteristics of the items in the test at an early stage of the calibration process. In addition, it is easier to detect fluctuations in the item parameters which may be caused by item exposure with the C strategy.

Research Questions

The main research question in this study was: Which of the three calibration strategies is the most effective for calibrating a new item bank effectively, while estimating ability precisely? In order to assess this in more detail the three strategies described above were compared based on a number of criteria: the global and conditional precision of ability estimates in a large calibration sample; the precision of the strategies at different points in the calibration process, and for different size calibration samples; the uniformity of item exposure; and their application under the assumptions of the 1-PL and 2-PL models. A secondary research question was: Could accounting for the uncertainty in the parameters in the calibration phase of a test improve the precision of the ability estimates? Both questions were evaluated with simulation studies.

Simulation Studies

To investigate which of the considered calibration strategies leads to the lowest overall mean absolute error (MAE) in the estimation of ability, simulation studies were conducted. Simulation studies make it possible to determine the true ability level of the test taker; next the calibration design can be reproduced in order to investigate the precision of the test result for each test taker. This cannot be done with operational data, because in practice it is impossible to assess the actual accuracy of a test since it is not possible to know the real ability level of a test taker. The simulation studies were programmed in Digital Visual Fortran 6.0 standard for Windows. The simulations 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 1-PL model and the 2-PL model.
4. The size of the item bank, varied as *K* = 100, 200, 400 items.
5. Accounting for uncertainty in the parameter estimates.

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. The length of the test was also varied as: 20, 30 and 40 items in certain conditions, however, only the results of the test with 20 items are reported in this paper in order to save space. The test length and item bank sizes selected for this simulation are typical for an occupational testing or certification program that uses a test battery with several unidimensional CAT's.

Method

The three calibration strategies were compared by assessing the accuracy of the ability estimate while in the calibration phase of the test. Once the number of test takers 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 test takers. A calibration sample of 4,000 test takers was set in this study.

The test takers' _ 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 1-PL or 2-PL 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**.** MAE was computed as the mean absolute difference between the true ability drawn from the N (0,1) distribution and the ability estimated by the weighted maximum likelihood procedure. The MAE for each strategy was then calculated by averaging across all test takers 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 test taker's score was estimated. This conditional precision was measured at specific points on the ability continuum (_ = -2, -1, 0, 1, 2), to give an estimate of the precision with which a test taker with a specific _ could be expected to be assessed within each condition. Therefore, after each phase 4,000 test takers were simulated at each of the five ability values, and the MAE was computed for each of the five ability values.

Results

Global Precision and Optimal Transition Points

Before the research questions could be investigated, it was necessary to identify 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 1-PL and 2-PL models. The results are shown in Table 2.

-- insert Table 2 here--

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 1-PL and 2-PL models. A comparison of the MAE for the three strategies indicated that the C strategy consistently resulted in the best ability estimates across all conditions.

The results for the 1-PL 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 2-PL model were similar to those for the 1-PL; 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 algorithm to administer items, assuming that the item parameters were accurate, resulted in inaccurate ability 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 1-PL and 2-PL models, in order to have comparable results across settings.

Local Comparison of the Calibration Strategies

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

--- insert figure 1 here---

Figure 1 illustrates the conditional precision of the three strategies with the 1-PL model on the left side and the 2-PL model on the right side, for item bank sizes of 100, 200, and 400 items. The horizontal axis represents the ability level of the test takers at five points on the _ scale (-2, -1, 0, 1, 2), and the vertical axis represents the MAE across the first 4,000 test takers within each design. For the 1-PL model, the graph shows that the C strategy measured ability more precisely than the other strategies at extreme ability scores, while all three strategies performed fairly equally at _ = 0. The use of random item administration with item parameter estimates of b_i = 0 performed well at _ = 0; however, this method performed much poorer at extreme ability levels. For the 2-PL 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 2-PL model, with the largest differences occurring at extreme ability 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 test takers. 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. Figure 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 ability for the nth test taker 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 nth test taker 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).

--- insert figure 2 here---

The results indicate that strategy C performed nearly as well as a fully calibrated test after as few as 500 test takers for the 1-PL model; it took strategy M 1,000 test takers 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 2-PL model were similar to the 1-PL 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 test taker at a particular point in the calibration process. Figures 2c and 2d present the cumulative precision of each strategy, which is the average precision with which a test taker 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 test takers in the calibration sample on the horizontal axis. The results were similar for the 1-PL and 2-PL 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 test takers in the calibration sample reached 500 for the 1-PL model, and after as few as 250 for the 2-PL model. The M and P2 strategies resulted in ability 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 test takers.

Item Exposure

The calibration strategies have been compared in terms of how accurately ability is assessed in the calibration process. However, the calibration strategy should also 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 test takers. The results for the 1-PL model are presented in the upper portion, and the 2-PL model in the lower portion of the table.

-- insert Table 3 here--

Table 3 shows a fairly uniform administration of items for all three calibration strategies for the 1-PL model. Item administration for the 2-PL 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.

Accounting for Uncertainty in the Parameter Estimates

In IRT item parameters are usually estimated, and then these estimates are treated as true parameters in subsequent analyses. Most of the literature on IRT takes this assumption for granted. However van der Linden and Glas (2000) discovered that the impact of estimation error can have dramatic consequences on ability estimation.

In the current study there is known uncertainty in the parameters, because ability is estimated with items that are in the calibration process. Therefore it is important to investigate the consequences of taking uncertainty into account in the model. Uncertainty can be taken into account by using a distribution of the parameter instead of a point estimate in the estimation equation. The distribution is simply the likelihood distribution associated with the parameter estimate, which represents the current level of confidence related to each parameter. Here four different conditions were assessed.

1. All item parameters were treated as true parameters.
2. Uncertainty in theta was taken into account in the model, but uncertainty in the item parameters was ignored.
3. Uncertainty in the item parameters was taken into account in the model, but uncertainty in theta was ignored.
4. Uncertainty in all parameters was taken into account in the model.

-- insert Table 4 here--

Table 4 presents the MAE for theta calculated under each of the four conditions for the 1-PL and 2-PL models for a calibration sample of 4,000. In general, taking uncertainty into account in all parameters decreased precision in theta estimation. The results also indicate that the use of a point estimate is better than a distribution in estimating theta. Taking uncertainty in the item parameters into account did not decrease precision greatly and slightly increases precision in the P2 strategy with the 1-PL model, and in the C strategy with the 2-PL model.

Discussion

The purposes of this paper was to investigate three different calibration strategies for calibrating an item bank from scratch, with the primary objectives of calibrating items in a fair and effective manner, while providing accurate ability estimates throughout the calibration design. The benefits of the three strategies were tested in terms of several possible conditions.

The C strategy consistently outperformed the other two strategies across all test lengths, and all item bank sizes. An example is that ability was estimated nearly as well as in a fully calibrated test after as few as 500 test takers in a test consisting of 20 items and an item bank consisting of 100 items for the 1-PL model. A weakness of this strategy was the non-uniform administration of items with the 2-PL 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 2-PL is used, because this strategy resulted in a more uniform administration of items with both models. However, a larger number of test takers were required before the precision in ability 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 ability estimates for test takers with ability estimates near the mean; however, this method was inaccurate at estimating test takers with extreme ability 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 test takers 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 2-PL model, because the item selection algorithm in the 2-PL 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 2-PL model, which leads to the selection of items with greater information at specific points on the ability scale, over items that provide information across a broader area. This can be efficient when there is little error in ability and item parameter estimates; however, it is not optimal at the beginning of a test when there is a lot of insecurity concerning a test taker’s ability, and is undesirable when there is error in the item parameters. The use of the 2-PL 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 2-PL 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 2-PL 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 2-PL 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 ability estimation using the 2-PL model with a fully calibrated test. They highlighted four solutions for controlling the capitalization of error in ability estimation: cross validation, controlling the composition of the item pool, imposing constraints, and using the 1-PL 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 1-PL model all three calibration strategies resulted in improved ability estimates, in addition to a uniform calibration of items. The results suggest that the 1-PL 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 2-PL model.

The study also investigated the consequences of accounting for the uncertainty in the parameter estimates with the 1-PL model. Accounting for this uncertainty lead to lower precision in most contexts, and a slight increase in others. The mixed results and the extra calculation time needed to account for the uncertainty in the parameters suggests that a point estimate would be preferred in most settings, even though there is possible error in the parameter estimates in the calibration phases of a test.

The results of the study provide viable calibration design options for test development orgnaizations that find it difficult to attract 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 in that they offer the possibility to assess test takers' ability throughout the calibration of the test. This makes it more attractive for test users and companies that purchase tests to become involved in the development phases of the test because the results can be used. It is important for practitioners to be aware of the ethical and legal consequences of administering scores while the test is in the calibration phase. Therefore, it is vital to have clear guidelines about how the results should be used at different points in the calibration process.

The cost of developing a CAT compared to its benefits will always be compared to other test designs. It is considerably more expensive to develop a CAT compared to a linear test. However, the long term benefits of a CAT may outweigh the initial costs, because items can be used longer, since they are exposed less frequently in this format. A cost-benefit analysis based on the expected item exposure, and the benefits of CAT for the specific testing program, can be conducted before a decision to develop a CAT is made.

Future research could investigate the consequences of using the 2-PL 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. In addition accounting for uncertainty in parameters did not increase accuracy greatly in this study; however Bayesian methods could be explored in future studies to investigate if these models can lead to better ability estimates when accounting for uncertainty. In addition these models can be used to incorporate pre-existing hypotheses about item parameters. 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 (2004)

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.

Table 1

M Strategy Design

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

Table 2

Comparison of the MAE for Different Transition Points within each Calibration Strategy

Model	Item bank	Strategy	MAE					
				T = 10	T = 25	T = 50	T = 100	T = 200
1-PL	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					
2-PL	K = 100	R	0.405					
		P2		0.475	0.388	0.353	**0.352**	0.361
		M		0.362	0.366	**0.349**	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 printed in bold.

Table 3

Number of Times Items were Administered for each Strategy within each Model

Model	Item bank	Strategy	Number of administrations						
			<100	100-199	200-399	400-599	600-799	800-999	>1000
1-PL	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
2-PL	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 4

Global Precision: MAE for Four Methods of Calculating Theta Based on Taking Uncertainty in Parameters into Account

Model	Strategy	All parameters treated as true parameters	Uncertainty in theta is taken into account	Uncertainty in the item parameters is taken into account	Uncertainty in all parameters is taken into account
1-PL	P2	.392	.394	.388	.390
	M	.389	.400	.392	.394
	C	.379	.392	.381	.388
2-PL	P2	.352	.366	.353	.367
	M	.349	.356	.354	.354
	C	.345	.350	.344	.359

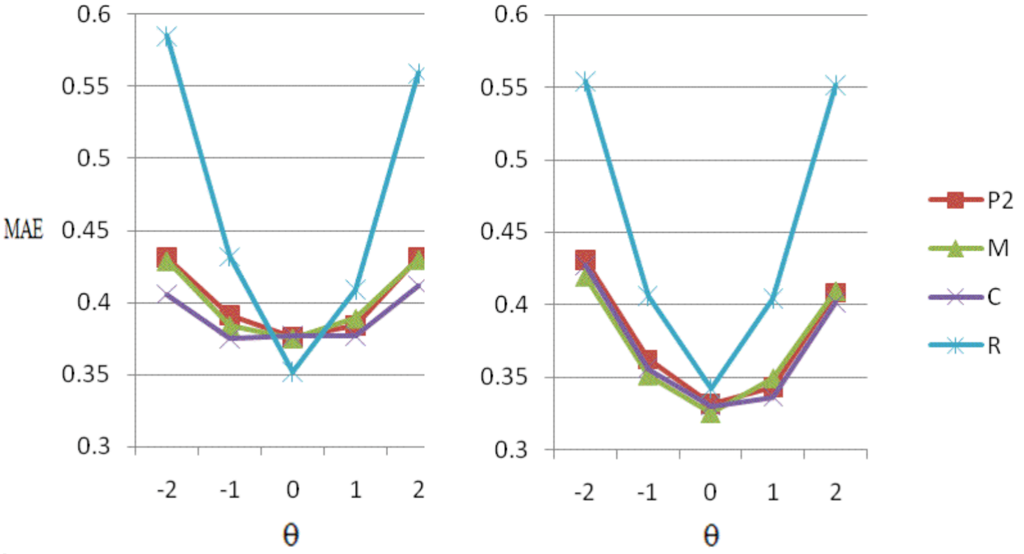
Figure Caption

Figure 1: Conditional precision: MAE at specific points on the _ continuum for strategies P2, M, C and random item administration (R) for the 1-PL model on the left, and the 2-PL model on the right, presented for an item bank size of 100 on the top, followed by an item bank size of 200, and finally an item bank size of 400 on the bottom.

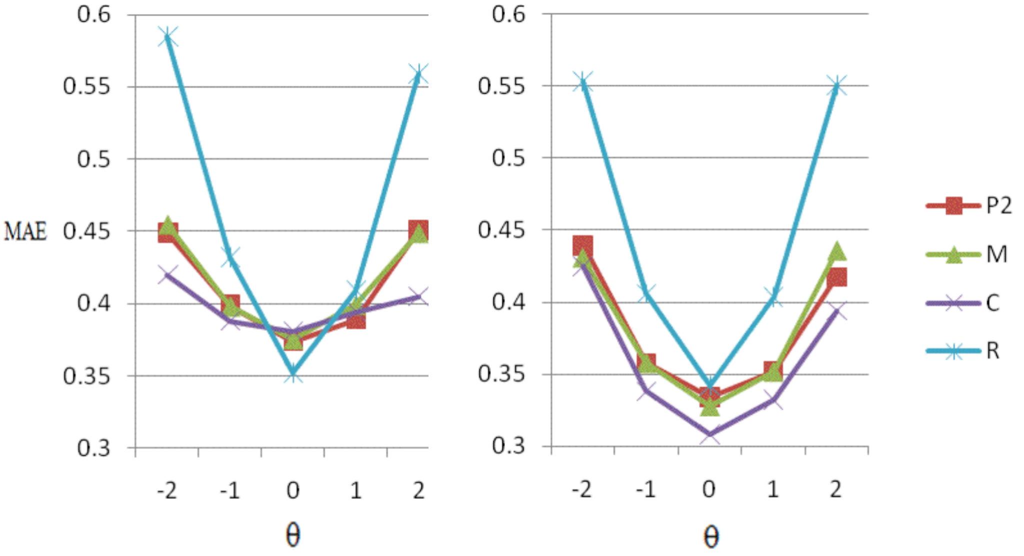
Figure 2: Top: Specific precision of each strategy, random item administration (R), and a completely calibrated test (O) for test taker number 250, 500, 1000, 2000, 3000 and 4000 for the 1-PL model on the left, and the 2-PL model on the right.

Bottom: Cumulative precision of each strategy and random item administration (R), for 250, 500, 1000, 2000, 3000 and 4000 test takers for the 1-PL model on the left, and the 2-PL model on the right.

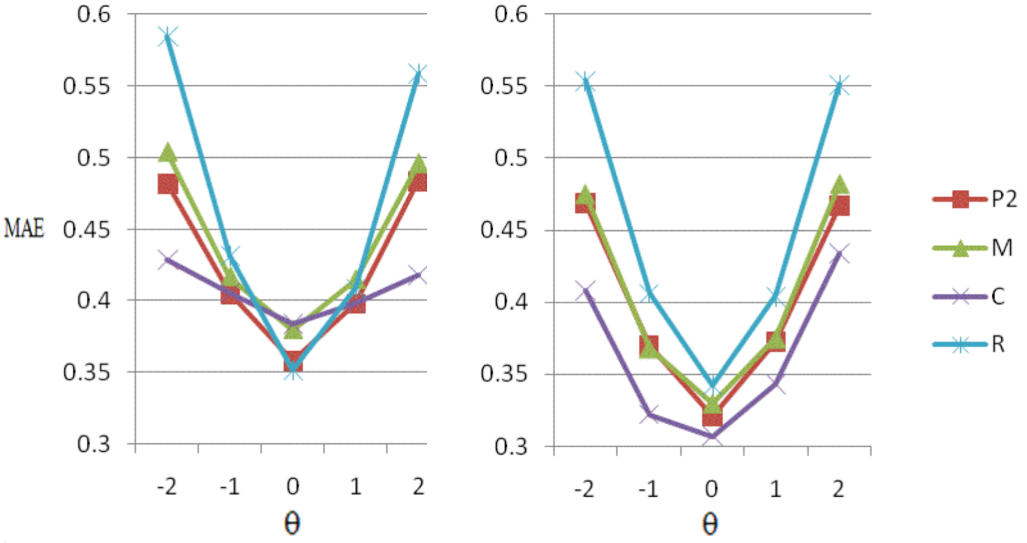
K = 100

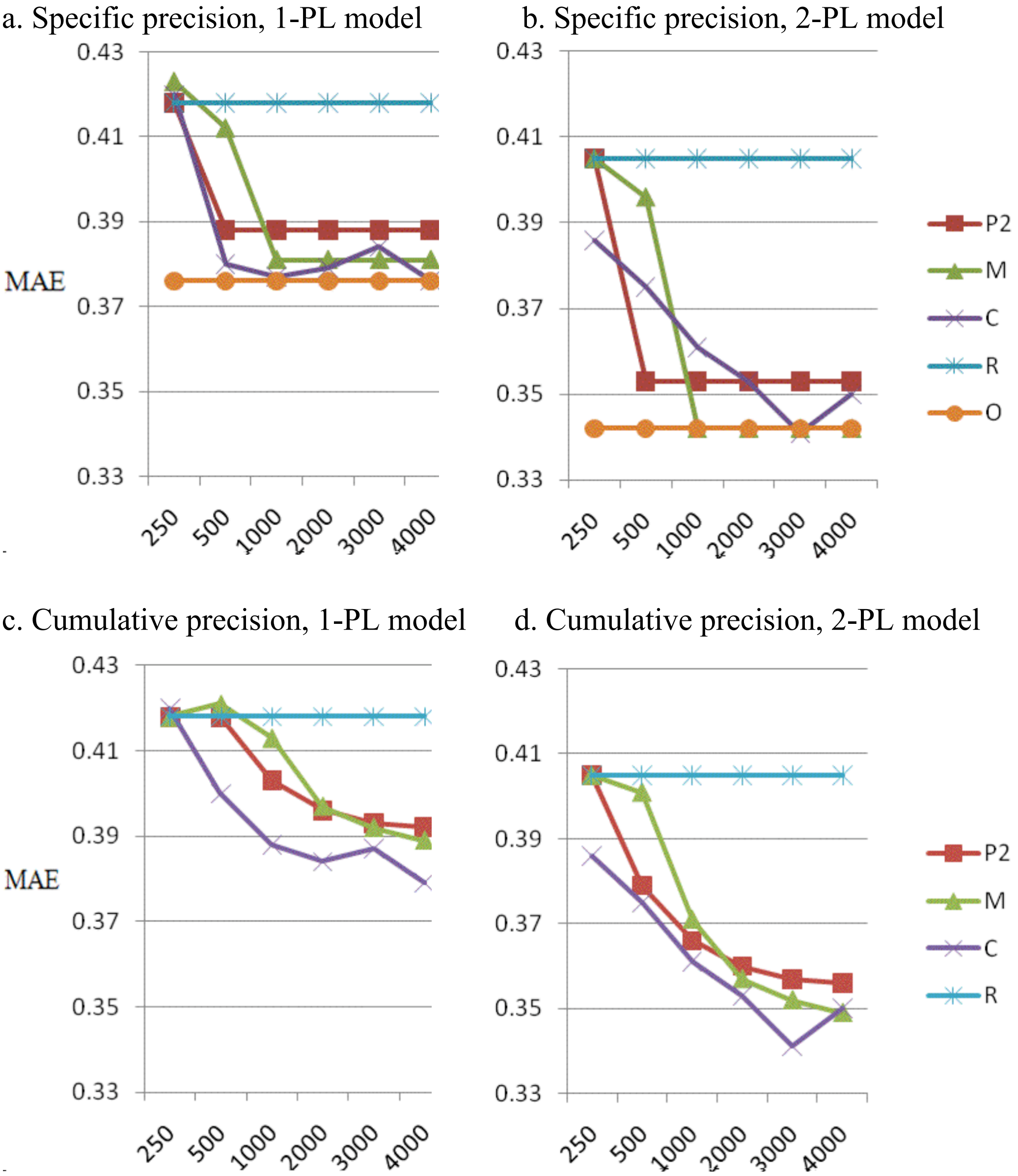


K = 200



K = 400





a. Specific precision, 1-PL model

b. Specific precision, 2-PL model

c. Cumulative precision, 1-PL model

d. Cumulative precision, 2-PL model