ISSN 1303-0485 • **eISSN** 2148-7561
DOI 10.12738/estp.2015.6.0102
Copyright © 2015 EDAM • http://www.estp.com.tr
Educational Sciences: Theory & Practice • 2015 December • 15(6) • 1585-1595

Received | June 2, 2015
Accepted | August 6, 2015
OnlineFirst | November 18, 2015

Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing*

Alper Şahin[a]
Çankaya University

David J. Weiss[b]
University of Minnesota

Abstract
This study aimed to investigate the effects of calibration sample size and item bank size on examinee ability estimation in computerized adaptive testing (CAT). For this purpose, a 500-item bank pre-calibrated using the three-parameter logistic model with 10,000 examinees was simulated. Calibration samples of varying sizes (150, 250, 350, 500, 750, 1,000, 2,000, 3,000, and 5,000) were selected from the parent sample, and item banks that represented small (100) and medium size (200 and 300) banks were drawn from the 500-item bank. Items in these banks were recalibrated using the drawn samples, and their estimated parameters were used in post-hoc simulations to re-estimate ability parameters for the simulated 10,000 examinees. The findings showed that ability estimates in CAT are robust against fluctuations in item parameter estimation and that accurate ability parameter estimates can be obtained with a calibration sample of 150 examinees. Moreover, a 200-item bank pre-calibrated with as few as 150 examinees can be used for some purposes in CAT as long as it has sufficient information at targeted ability levels.

Keywords: Computerized adaptive testing • Calibration sample size • Ability estimation accuracy • Pretest item calibration • Item Response Theory

* This study was funded by The Scientific and Technological Research Council of Turkey (TUBITAK) within the International post-doctoral research fellowship programme with fund number 1059B191300180 / An earlier version of this paper was presented at the International Congress on Education for the Future: Issues and Challenges, Ankara, May 2015.

a **Corresponding author**
Alper Şahin (PhD), Academic English Unit, Çankaya University, Mimar Sinan Caddesi No: 4, Ankara 06790 Turkey
Research areas: Item Response Theory; Computerized adaptive testing; Language testing; Calibration sample size; Item and person parameter estimation
Email: alpersahin2@yahoo.com

b Prof. David J. Weiss (PhD), Department of Psychology, University of Minnesota, 75 E River Rd, Minneapolis, MN 55455 U.S.
Research areas: Computerized adaptive testing; Item Response Theory; Adaptive measurement of intra-individual change
Email: djweiss@umn.edu

Nearly 50 years of technical research and recent developments in computer technology have made computerized adaptive testing (CAT) applications more feasible and affordable for educational institutions worldwide. There are numerous advantages of using a CAT platform to deliver tests: (*i*) CAT requires less testing time, (*ii*) the test result can be calculated immediately, (*iii*) the test is easier to deploy and less vulnerable to theft, and (*iv*) it can be administered wherever and whenever needed (Hambleton & Swaminathan, 1985; Rudner, 1998).

The success of a CAT program highly depends on a large item bank, which is maintained regularly, with items distributed across a wide range of ability (θ) levels. Such an item bank is necessary to obtain accurate θ estimates for examinees whose latent trait will be estimated. However, the preparation of such a bank entails some challenges. One challenge, possibly the most important, is that items that will be placed in a CAT item bank must be pretested and calibrated on the same scale. Highly accurate item parameters are desired because θ estimates in CAT applications are based on these parameters. A critical variable confounds item parameter estimation at this stage: the size of the examinee sample that will be used to pretest items in the bank.

Sample Size Requirements in Item Response Theory-Based Item Calibration

The item parameter calibration process for a CAT item bank is conducted using item response theory (IRT) models. IRT typically requires large sample sizes for accurate item parameter estimation (Hambleton, 1989). This is largely based on a previous study by Lord (1968), who concluded that the standard errors of item discrimination parameters were very high until a test of 50 items and a sample of 1,000 examinees was used in the three-parameter logistic model (3PLM), and later studies that were concerned with the calibration sample size also supported Lord's finding. Swaminathan and Gifford (1979) found that a sample of 1,000 examinees was necessary to estimate item parameters with high accuracy in the 3PLM. Hulin, Lissak, and Drasgow (1982) also concluded that a sample of 1,000 was necessary with 60 items to accurately estimate item parameters in the 3PLM. Although Ree and Jensen (1980) stated that accurate item parameter estimates require only 500 examinees in the 3PLM, with empirical support from studies by Patsula and Gessaroli (1995); Tang, Way, and Carey (1993); Yen (1987); and Yoes (1995), Lord's (1968) suggestion to use 1,000 examinees as the minimum item calibration sample size was accepted by many IRT researchers. However, some studies that supported Ree and Jensen's finding that sample sizes less than 1,000 can be used without losing much estimation accuracy were also published. A study conducted by Gao and Chen (2005) found that an item calibration sample of 500 can be used to accurately estimate item parameters in the 3PLM. Moreover, Weiss and von Minden (2012) obtained accurate item parameter estimates with a calibration sample of 200 examinees in the 3PLM. Finally, Akour and Al Omari (2013) found that a sample size of 500 was adequate to accurately estimate item parameters with 30 items in the 3PLM. It was somewhat expected that better results can be obtained with a sample size of 500 after 1995 because more advanced parameter estimation procedures were being used (e.g., marginal maximum likelihood; Baker & Kim, 2004) compared with those used in 1968. However, these studies could not gain sufficient support from practitioners, and Lord's (1968) suggestion to use 1,000 examinees to estimate item parameters is widely followed even today.

Calibration Sample Sizes and θ Estimation in CAT

Although IRT-based calibration sample size studies that focus on item parameter recovery have implications for the sizes of samples to be used in CAT pre-test item calibrations, it would be more useful to determine the effects of calibration sample size on θ estimation accuracy in a CAT environment. Surprisingly, there appears to be only two studies on calibration sample size and its effects on θ estimation in CAT.

Ree (1981) conducted a simulation study of calibration sample size in adaptive testing, in which sample sizes of 500, 1,000, and 2,000 examinees and item banks with 100, 200, and 300 items were simulated. He calibrated the items in the banks with different sample sizes and estimated θ in fixed-length CAT of 10, 15, 20, 25, 30, and 35 items. High correlations between true θ and estimated θ levels were observed when 20 or more items were administered. In addition, Ree concluded that a 200-item bank calibrated with 2,000 examinees is required to reduce the absolute error of θ estimation to acceptable levels in a CAT environment.

Chuah, Drasgow, and Luecht (2006) studied item parameter estimation accuracy for θ estimation in computerized adaptive sequential tests (CAST) in a simulation study and found that items pre-calibrated

with 300 examinees using the 3PLM can be used to accurately estimate examinee θ and to classify the examinees as masters or non-masters using CAST.

Purpose

Calibration sample size for CAT programs has implications not only for item parameter estimation accuracy but also for the start-up and maintenance costs of CAT programs. Although it might not constitute a serious problem for a test publisher with nearly unlimited resources to obtain large examinee samples, it might not be a feasible option for educational researchers who are working for institutions in developing countries. It is arduous for these researchers to obtain large examinee groups that embody the characteristics of target examinees to pretest items, and there has been little research on calibration sample size and θ estimation in CAT. Therefore, there is a need for a study that investigates the feasibility of using small sample sizes to calibrate items in a bank and the effects of calibration sample size on θ estimation in CAT. The present study was designed to satisfy the abovementioned need. Moreover, because the number of items in the CAT item bank is another potential source of high cost in CAT development, the effects of calibration sample size on θ estimates were investigated conditional on bank size. For this purpose, an answer was sought to the question "how do θ estimates based on item parameter estimates obtained from varying sample sizes and bank sizes recover the θ estimates from a large item bank calibrated using a very large sample of examinees?"

Method

Research Data Generation

The full dataset of the present study was simulated using the 3PLM and a Monte-Carlo simulation procedure in CATSim software version 4.0.6 (Weiss & Guyer, 2012a) with a uniform distribution of θ parameters between −3 and +3. The 3PLM is

$$P_{ij}(\theta_j) = c_i + (1-c_i)\frac{\exp\left[Da_i(\theta_j - b_i)\right]}{1+\exp\left[Da_i(\theta_j - b_i)\right]}, i = 1\ldots, n, \qquad (1)$$

where $P_{ij}(\theta_j)$ is the probability of a correct response to item i conditional on θ for examinee j, a_i is item discrimination, b_i is item difficulty, and c_i is the pseudo-chance parameter estimated for item i. The 3PLM considers the chance or guessing parameter; it was used in this study because multiple-choice items are frequently used in schools, and there is a certain degree of probability of giving a correct response to an item by chance in this question type. Thus, a model that ignores this chance/guessing variable would not have been appropriate for tests that use multiple-choice items.

All item parameters were generated from uniform distributions: a parameters ranged from 0.5 to 1.5, b parameters ranged from −3 to +3, and c parameters ranged from 0.00 to 0.25. As a result, a dataset of 10,000 examinees and 500 items was obtained. This dataset was designed to reflect an operational CAT with a 500-item bank that was pre-calibrated using 10,000 examinees.

Item Selection for Banks of Different Sizes

From the full bank with 500 items, items for the medium (300 and 200 items) and small (100 items) banks were selected using a systematic approach to maintain the same quality across all banks. Using this approach, from the simulated 500-item bank (Bank A), a sample of 400 items was drawn using the a parameters as strata in SPSS 20's (IBM Corporation, 2011) complex samples module. Then, another 400-item sample was drawn from the 500-item bank using the b parameters as strata. In this manner, two item samples that reflected the distributions of the a and b parameters in the 500-item bank were obtained. From these item sets, items that were common in both samples were taken into the final sample of items. There were 310 items common in both samples of 400 items. To resolve this problem, the items were ranked according to their a parameters first and then according to their b parameters. Ten pairs of items with very similar a and b parameters were identified, and the items with higher c parameters were eliminated from the sample, resulting in the 300-item bank (Bank B). To select items for the 200-item bank, two sets of 300 items were drawn from the 500-item bank using the a and b parameters as strata. Items that were common in both sets were taken into the final set of items. Item pairs with similar a and b parameters were identified, and those with higher c parameters were eliminated, resulting in the 200-item bank (Bank C). For the 100-item bank (Bank D), two sets of 200 items were drawn from the 500 items taking a and b parameters as strata, and the same procedure that was used for selecting the items in the 300- and 200-item banks was followed.

During the item selection procedure for 100-, 200- and 300-item banks, the responses of the simulated 10,000 examinees to the selected items

for smaller banks were kept intact and transferred to form datasets of 300, 200, and 100 items. In this manner, four datasets with 500 (full set), 300, 200, and 100 items and 10,000 examinee responses were obtained. The drawing of the examinee samples that were used to recalibrate items in the 100-, 200- and 300-item banks was done based on the datasets that had the original (simulated) responses of the simulated examinees to the items selected.

Drawing Calibration Samples

To have the examinee θ distribution in the full dataset reflected in the drawn samples, the examinees' θ levels were converted into categorical data by assigning a category number to θs at interval of 0.25 (e.g., θ = 3.00…2.75 = 1; θ = 2.749…2.50 = 2); in this manner, 24 discrete θ levels were obtained. Then, using the θ levels as strata in SPSS 20's (IBM Corp., 2011) complex samples module, samples of 150 (Harwell & Janosky, 1991), 250 (Goldman & Raju, 1986; Harwell & Janosky, 1991), 500 (Akour & Al-Omari, 2013; Baker, 1998; Gao & Chen, 2005; Goldman & Raju, 1986; Hulin et al., 1982; Thissen & Wainer, 1982), 1,000 (Goldman & Raju, 1986; Hulin et al., 1982; Lord, 1968; Thissen & Wainer, 1982; Weiss & von Minden, 2012; Yen, 1987), 2,000 (Gao & Chen, 2005; Hulin et al., 1982; Ree & Jensen, 1980; Yoes, 1995), 3,000 (Tang et al., 1993), and 5,000 (Akour & Al-Omari, 2013) that had been tested in previous research (including those conducted in one- and two-parameter logistic models) on IRT-based calibration sample size as well as two uncommon sample sizes (350 and 750) were drawn. These samples were drawn from each of the datasets with 100, 200, 300, and 500 items and 10,000 examinee responses. Therefore, 40 datasets (36 calibration samples and 4 full datasets) were obtained, as summarized in Table 1.

Item Calibration and θ Estimation Through Post-Hoc Simulations

After item selection and sampling, the 3PLM parameters of the items in the 36 sample datasets were re-estimated with marginal maximum likelihood estimation (MMLE) using default options in Xcalibre 4.2 (Guyer & Thompson, 2011). The estimated item parameters obtained from the samples were treated as the known item parameters in post-hoc simulations performed in CATSim (Weiss & Guyer, 2012a). Post-hoc simulations function as the last step before a live CAT administration. They are used to evaluate the CAT item bank, giving the CAT developer the opportunity to manipulate various parameters before a live CAT so that optimal CAT application procedures can be obtained. A post-hoc simulation requires a matrix of examinee responses to items in a CAT item bank and item parameters that are known for items in the bank. The simulation utilizes the examinee responses to simulate how the CAT item bank would function if the examinees actually faced items in banks in a live CAT (Weiss & Guyer, 2012b).

In the present study, full datasets with 10,000 examinees and 500, 300, 200, and 100 items and the item parameters calibrated using 36 samples were used in the post-hoc simulations. Thus, 36 CAT simulations were performed, one for each combination of sample size and bank size. In these simulations, 0.0 was used as the initial θ estimate for all examinees. Bayesian estimation by maximum a posteriori was used with a mean of 0.0 and standard deviation of 1.0. Maximum information at the estimated θ level was used as the item selection rule, and the CAT was terminated when the standard error of the θ estimate was 0.20 or less or when all items in the bank had been used. As a result of the post-hoc simulations, 36 CAT θ estimates were obtained for each person in the 10,000-examinee pool.

Table 1
Item Banks and Samples Drawn.

	Bank A (Simulated)	Bank B (Sampled)	Bank C (Sampled)	Bank D (Sampled)
Number of items	500	300	200	100
Number of examinees	10,000	10,000	10,000	10,000
Number of calibration samples drawn from the bank	9 (5,000 × 500, 3,000 × 500, 2,000 × 500, 1,000 × 500, 750 × 500, 500 × 500, 350 × 500, 250 × 500, 150 × 500)	9 (5,000 × 300, 3,000 × 300, 2,000 × 300, 1,000 × 300, 750 × 300, 500 × 300, 350 × 300, 250 × 300, 150 × 300)	9 (5,000 × 200, 3,000 × 200, 2,000 × 200, 1,000 × 200, 750 × 200, 500 × 200, 350 × 200, 250 × 200, 150 × 200)	9 (5,000 × 100, 3,000 × 100, 2,000 × 100, 1,000 × 100, 750 × 100, 500 × 100, 350 × 100, 250 × 100, 150 × 100)
Total number of datasets	10	10	10	10

θ Levels Treated as True θ

To obtain the "true" θ levels of the simulated 10,000 examinees, first, item parameters for the simulated full dataset (10,000 × 500) were estimated, and a post-hoc simulation was administered in CATSim using these estimated parameters. Thereafter, the θ range of the 10,000 individuals in the full dataset was found to be between −2.35 and +2.05. Estimated item parameters of the full dataset were used in this process because all item parameters in all research conditions were estimated. Thus, the possible discrepancy between the true and estimated parameters of the full dataset that were attributable to the estimation error caused by the parameter estimation software was eliminated. The θ estimates obtained after this last simulation were taken as the true θ levels of the simulated examinees (Swaminathan, Hambleton, Sireci, Xing, & Rizavi, 2003) and they were compared with those obtained after the 36 post-hoc simulations.

Evaluation of Estimation Accuracy

To evaluate estimation accuracy, correlations (Gao & Chen, 2005; Harwell & Janosky, 1991; Hulin et al., 1982; Yen, 1987) between the CAT θ estimates that were obtained after the 36 simulations and the true θ levels were calculated. Moreover, root-mean-squared difference (RMSD) (Gao & Chen, 2005; Harwell & Janosky, 1991; Yen, 1987) and average signed difference (ASD) were also calculated for these θ estimates using Equations 2 and 3:

$$RMSD\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \hat{\theta}_{Ti})^2}{N}}, \quad (2)$$

$$ASD\left(\hat{\theta}\right) = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \hat{\theta}_{Ti})}{N}, \quad (3)$$

where θ_j represents the estimated θ level for the jth examinee for each research condition tested, θ_{Ti} represents the true θ level for each examinee as defined above and N is the number of examinees.

Bank Information Functions

Bank information functions (BIFs) indicate how well examinees' θ levels at a specific θ level would be measured if all items in an item bank were used to estimate θs. Moreover, the amount of information obtained from an item bank at a specific θ level is inversely related to the conditional standard error of measurement (Hambleton & Swaminathan, 1985). BIFs pertaining to the 500-, 300-, 200-, and 100-item banks were plotted after item parameters in the full simulated dataset were estimated. Figure 1 indicates that item banks obtained with estimated item parameters had similar BIFs that covered similar θ levels as desired. Moreover, the highest information is obtained around $\theta = 1$, and the lowest information level is on both sides of the θ continuum.

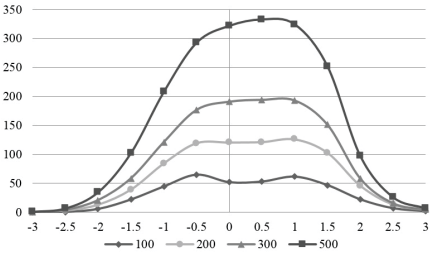


Figure 1: BIFs for 100-, 200-, 300-, and 500-item banks.

Results

Correlations between the θ estimates that were obtained with the 500-, 300-, 200-, and 100-item banks and sample sizes varying from 150 to 5,000 are presented in Figure 2. As shown in the figure, the correlations were all over 0.94 regardless of the sample size used to calibrate items and the bank size employed. Although there was a slight increase in the correlations across the sample sizes that were used to calibrate items, the correlations obtained ranged within a very narrow interval, roughly between 0.94 and 0.98. Such high correlations indicate strong positive linear relations between the true θ and estimated θ regardless of bank size or number of examinees. Although correlations remained essentially constant or increased with sample size for most bank sizes, they slightly decreased as the sample size increased in the 100-item bank.

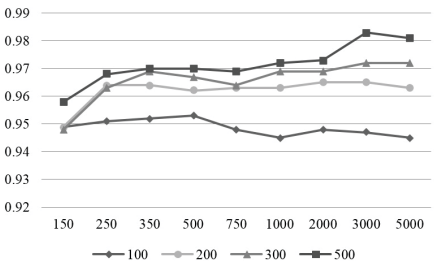


Figure 2: Correlations between θ estimates for banks with 100, 200, 300, and 500 items and sample sizes of 150–5,000.

The RMSDs and ASDs for the θ estimates are presented in Figures 3 and 4. As can be seen, there was a decreasing trend in RMSDs (Figure 3) as the calibration sample size increased (except for the 100-item bank), as was expected. Moreover, the increase in the item bank size from 100 to 500 items resulted in a decrease in RMSDs. Differences in RMSDs were minimal among the 200-, 300-, and 500-item banks for sample sizes of 2,000 or less and slightly larger for the two largest sample sizes.

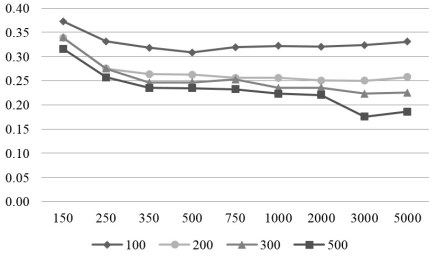


Figure 3: RMSDs for banks with 100, 200, 300, and 500 items and sample sizes of 150–5,000.

The ASDs shown in Figure 4 fluctuated when the sample size increased from 150 to 350 and stabilized with sample sizes of 500 or larger. After this size, ASDs remained around 0.0 for all sample sizes across different bank sizes (except for the 100-item bank). Although larger ASDs were obtained for the 100-item bank, the results still indicated a very low (less than −0.05) amount of negatively biased estimation of θ for sample sizes of 350 and larger.

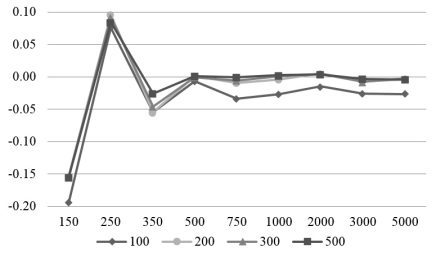


Figure 4: ASDs for banks with 100, 200, 300, and 500 items and sample sizes of 150–5,000.

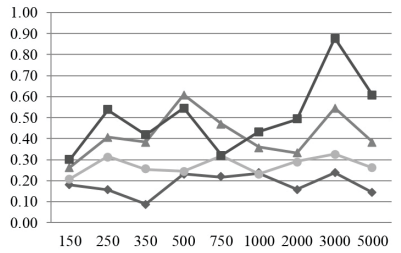
Correlations Conditional on θ Groups

Overall, Figures 2, 3, and 4 indicate that accurate θ estimation can be obtained across all sample sizes and item banks, as suggested by the very high correlations and low RMSDs and ASDs between the true θ and estimated θ. However, to determine the effects of the item bank and sample size interaction on θ estimation accuracy in CAT applications considering the banks' item information levels better, the correlations, RMSDs, and ASDs pertaining to these estimates were computed conditional on the θ continuum divided into five θ groups (Group 1, $\theta = -2.35$ to -2.00; Group 2, $\theta = -1.99$ to -1.00; Group 3, $\theta = -0.99$ to 0.00; Group 4, $\theta = 0.001$ to 0.99; Group 5, $\theta = 1.00$ to 2.05). The numbers of examinees (N) in each θ group were 42, 1,694, 3,012, 3,371, and 1,881 for Groups 1, 2, 3, 4, and 5, respectively.

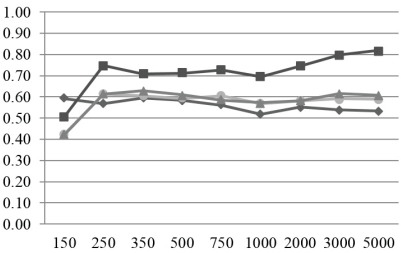
The correlations obtained conditional on the θ groups are shown in Figure 5; the θ estimates in Group 1 (Figure 5a) were not highly correlated with the true θ levels of the examinees in this group. Correlations were 0.60 or less under all conditions of the item bank size and sample size with the exception of the N = 3,000, 500-item bank condition, in which the correlation approached 0.90. For all sample sizes, the correlations were somewhat erratic, possibly because of the small number of examinees in this group.

The correlations obtained for the examinees with θ levels from −1.99 to −1.00 (Group 2, Figure 5b) were generally relatively higher and more stable. However, correlations above 0.70, which indicates a moderate correlation (Yoes, 1995), were obtained only after the item bank size increased to 500 and when the calibration sample size was 250. The correlations generally increased as the bank size increased from 100 to 500, but differences between the correlations were trivial among the 300-, 200-, and 100-item banks. For example, the correlations obtained from the 300-, 200-, and 100-item banks calibrated with 250 examinees were 0.614, 0.614, and 0.567, respectively. However, the correlation obtained from the 500-item bank in the same condition was 0.748.

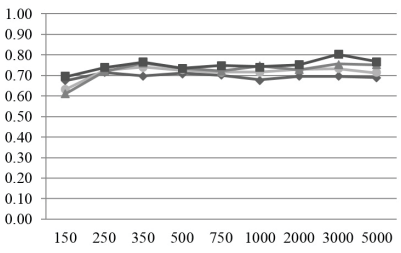
The correlations for Group 3 (Figure 5c) ranged between 0.611 and 0.803 across the calibration samples. The correlations for this θ range were mostly between 0.70 and 0.80, with a slight reduction as the bank size decreased in most cases. For example, the correlations obtained from items that were calibrated with 350 examinees in Group 3 were 0.764, 0.759, 0.740, and 0.697 for 500-, 300-, 200-, and 100-item banks, respectively. This decrease was also observed in samples of 750, 3,000, and 5,000 and partially observed in samples of 250, 500, 1,000, and 2,000. Figure 5c shows that the correlations for item banks of 300 and 500 items were very close to each other, with sometimes slightly higher correlations (N = 500 and 1,000) obtained with the 300-item bank. Slightly higher correlations were obtained as the calibration sample size increased.



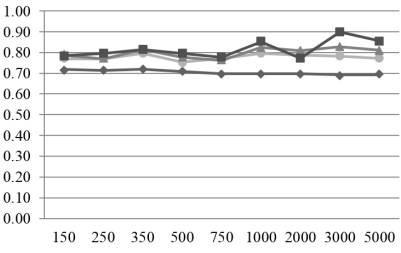
5a. Correlations for Group 1 (θ = −2.35 to −2.00, N = 42)



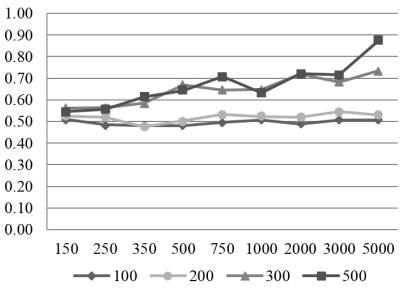
5b. Correlations for Group 2 (θ = −1.99 to −1.00, N = 1,694)



5c. Correlations for Group 3 (θ = −0.99 to 0.00, N = 3,012)



5d. Correlations for Group 4 (θ = 0.001 to 0.99, N = 3,371)



5e. Correlations for Group 5 (θ = 1.00 to 2.05, N = 1,881)

Figure 5: Correlations conditional on θ groups for banks with 100, 200, 300, and 500 items and samples of 150–5,000.

The correlations pertaining to Group 4 in Figure 5d ranged between 0.698 and 0.899 across calibration samples. Slightly lower correlations were obtained as the bank size decreased, as was the case in Group 3, with a larger decrease for the 100-item bank. The correlations for the 100-item bank were uniformly the lowest and were essentially constant across calibration samples. Quite similar correlations were obtained across the calibration samples for the other bank sizes with the exception of the 500-item bank with larger sample sizes when there was an increase for the sample sizes of 3,000 and 5,000. Higher correlations were obtained for Group 4 than for Groups 1, 2, and 3.

The correlations for Group 5 (high θ group) in Figure 5e ranged between 0.475 and 0.874. Note that the correlations obtained from the 300- and 500-item banks were quite similar and moved on quite similar trajectories across different calibration sample sizes. There was a linear relation between the sample size and estimation accuracy for these item banks. The correlations obtained from the 300- and 500-item banks were higher than those obtained from the 100- and 200-item banks for this θ range, which were essentially identical to each other.

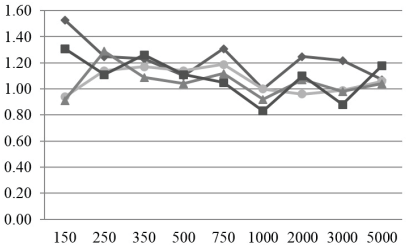
RMSDs and ASDs Conditional on θ Group

The RMSDs and ASDs that were calculated conditional on θ groups are presented in Figures 6 and 7. In contrast to the low RMSDs in Figure 3, the values for θ Group 1 (Figure 6a) were high, with average RMSDs of 1.22, 1.07, 1.05, and 1.09 in the 100-, 200-, 300-, and 500-item banks, respectively.

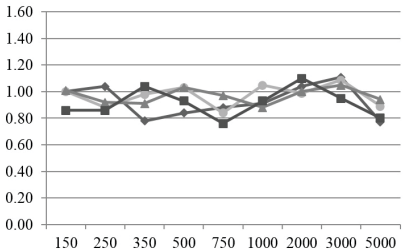
The RMSDs for Group 2 are presented in Figure 6b; they were lower than those obtained for Group 1, but were still high, with average RMSDs of 0.93, 0.97, 0.97, and 0.92 for the 100-, 200-, 300-, and 500-item banks, respectively. The correlations for this group were moderate in some cases for the 500-item bank.

The RMSDs obtained for Group 3 are presented in Figure 6c; a substantial decrease was observed. The values ranged between 0.37 and 0.67, lower than those obtained for Groups 1 and 2.

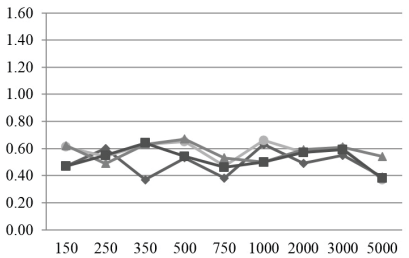
The RMSDs for Group 4 are presented in Figure 6d; they ranged between 0.25 and 0.38. It is clearly observed in the figure that highly similar RMSD values were obtained across all item banks in all sample sizes. Although the RMSDs tended to be lower as the sample size increased, the magnitude of the change was trivial.



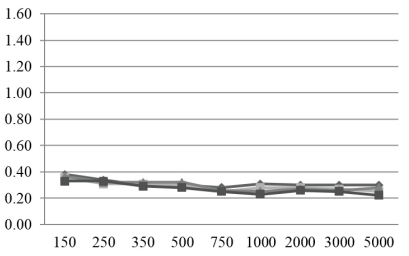
6a. RMSDs for Group 1 (θ = −2.35 to −2.00, N = 42)



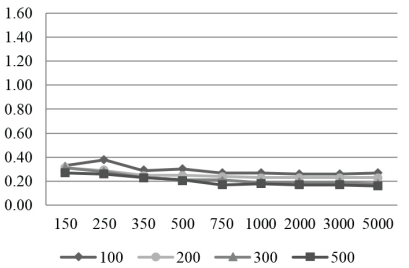
6b.RMSDs for Group 2 (θ = −1.99 to −1.00, N = 1,694)



6c. RMSDs for Group 3 (θ = −0.99 to 0.00, N = 3,012)

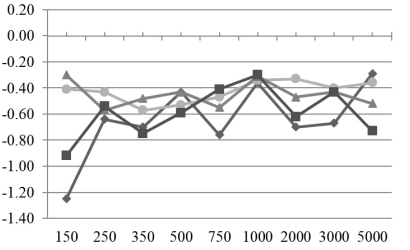


6d. RMSDs for Group 4 (θ = 0.001 to 0.99, N = 3,371)

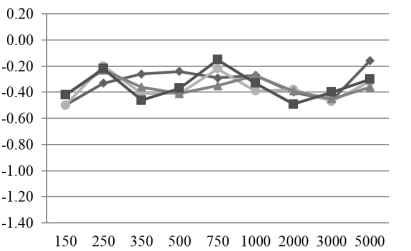


6e.RMSDs for Group 5 (θ = 1.00 to 2.05, N = 1,881)

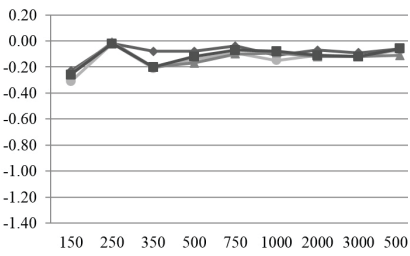
Figure 6: RMSDs conditional on θ groups for banks with 100, 200, 300, and 500 items and samples of 150–5,000.



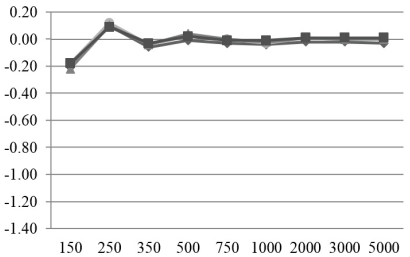
7a. ASDs for Group 1 (θ = −2.35 to −2.00, N = 42).



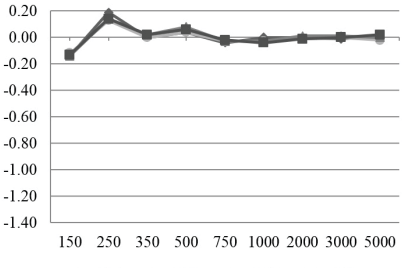
7b. ASDs for Group 2 (θ = −1.99 to −1.00, N = 1,694



7c. ASDs for Group 3 (θ = −0.99 to 0.00, N = 3,012)



7d. ASDs for Group 4 (θ = 0.001 to 0.99, N = 3,371)



7e. ASDs for Group 5 (θ = 1.00 to 2.05, N = 1,881)

Figure 7: ASDs conditional on θ groups for banks with 100, 200, 300, and 500 items and samples of 150–5,000.

The RMSDs obtained for Group 5 are presented in Figure 6e. As shown in the figure, the values ranged between 0.20 and 0.29, even lower than those obtained for Group 4. This indicates highly accurate θ estimation across all bank sizes and sample sizes.

The ASDs conditional on θ groups are presented in Figure 7. The θ estimates for Group 1 (Figure 7a) had negative ASDs with values ranging between −1.25 and −0.29. This indicates a large amount of negative bias in θ estimates for this range, indicating substantial underestimation of the true θs. The ASDs for Group 2 (Figure 7b) ranged between -0.15 and -0.50 and were high, indicating negatively biased estimates of θ within this range. The ASDs for Group 3 (Figure 7c) ranged between -0.01 and −0.31, which indicated low ASDs between the true θ and estimated θ. As shown in Figure 7c, the ASDs were essentially the same among item banks of different sizes across all sample sizes in this range. The ASDs obtained for Group 4 (Figure 7d) ranged between 0.12 and −0.22, indicating better θ estimation in this group than in the previous three groups. Finally, the ASDs for Group 5 (Figure 7e) ranged between −0.14 and 0.19, which also indicated biased but accurate θ estimates.

Discussion and Conclusions

The results for the overall correlations, RMSDs, and ASDs that were calculated for each item bank and sample size combination indicated that although sample size and bank size influenced θ estimation in CAT, this influence was negligible. However, when the results from the θ groups were analyzed, more detailed implications in terms of the effects of the item bank quality and sample size interaction on θ estimation in CAT were obtained.

The results for θ Group 1 (θ = −2.35 to −2.00) indicated that the bank quality was more important than the calibration sample size on examinee θ estimation in CAT. This was the only θ range for which item banks had uniformly low information levels and thus had the lowest capability of providing measurement accuracy. The results also indicated that if an examinee's θ and item bank information did not match, inaccurate θ estimates for examinees with those θ levels were obtained. A more important finding was that regardless of how large sample size was used to calibrate the items, it seemed to have no effect on improving the θ estimates when the bank lacked sufficient information for a particular θ range. This suggests that if the item bank does not have sufficient information for a particular θ range, calibration sample size cannot be used as a means to increase θ estimation accuracy in that range in CAT.

The item banks used in the present study had more information in θ Group 2 (θ = −1.99 to −1.00) compared with θ Group 1. However, the results for this group confirmed the findings in Group 1. Although the item banks had more information for this range, it was not sufficient to accurately estimate θ. The correlations, RMSDs, and ASDs indicated inaccurate θ estimates. This again was because of the lack of sufficient information in the banks for this θ range, as shown in Figure 1. However, the item bank size had a somewhat positive effect on the θ estimates in this range in that more items in the item bank resulted in more information. The results from Group 2 also confirmed the finding that the calibration sample size did not improve CAT θ estimates in θ ranges for which item banks had little information.

Similar correlations, RMSDs, and ASDs were obtained for examinees in Group 3 (θ = −0.99 to 0.00) across all item bank and calibration sample sizes. Note that the item bank size lost the influence on θ estimation, which was clearly observed in Groups 1 and 2. This was possibly because of the higher level of information that item banks had in Group 3, and it suggests that whether an examinee's θ falls into areas where an item bank has sufficient information will determine the magnitude of the effect of the item bank size on θ estimation accuracy.

The results from Group 4 (θ = 0.001 to 0.99) and Group 5 (θ = 1.00 to 2.05) confirmed the findings for Group 3; the correlations, RMSDs, and ASDs were very close to each other across all item banks and calibration sample sizes in these groups. It can be clearly seen that the information functions of the item banks illustrated in Figure 1 had their peaks at or around θ = 1.0. This means that the θ ranges of Groups 4 and 5 covered the area at which all item banks had their highest levels of information, which resulted in higher θ estimation accuracy, as indicated by the correlations, RMSDs, and ASDs. However, as in Group 3, the item bank and calibration sample sizes had almost no influence on the accuracy of θ estimates when there was more information in a bank that matched the target examinee's θ level.

The findings of the present study indicate that calibration sample size had little or no influence on θ estimation in CAT applications as long as the bank had sufficient levels of information in the regions of the banks where individual θs were located. In the present study, accurate θ parameter estimates

were obtained even in research conditions in which a sample of 150 examinees was used to calibrate the items in the banks. Moreover, a bank of 100 items could function as successfully as the banks with as many as 500 items, especially at certain θ levels, depending on the information that pertained to those θ levels in the banks. This suggests that a 100-item bank could serve for some purposes if it was developed using quality items that provided information at appropriate θ levels. However, because a bank's information level is highly dependent on the number of items in it, having a bank of 200 items or more would serve better for most purposes and would decrease the risk of having inaccurate θ estimates in most situations. Moreover, in most situations in the present study, the 300-item bank served as well as the 500-item bank. This might mean that if a large item bank is planned to be developed, a bank of 300 items could be considered.

It should also be kept in mind that these suggestions are valid as long as the item banks have sufficient information at the θ levels that match the target examinees. From that point of view, the item banks that were used in this study were not optimal for CAT, which requires a bank with an essentially horizontal information function for optimal performance. Different results likely would have been obtained conditional on θ for this type of bank. However, such banks are difficult to construct, and real item banks are likely to be somewhere in between these two extremes.

The findings of the present study were partially in parallel with the previous literature, in which there was a limited number of similar studies. An item bank of 200 items calibrated on 2,000 examinees was what Ree (1981) suggested as necessary to obtain accurate θ estimates using CAT. In the present study, a bank of 200 items was found to be feasible for some purposes as well. However, a calibration sample of 150 examinees was also found to be feasible for obtaining reasonably good θ estimates in situations in which the item bank had sufficient information for the target θ. This difference in results is partly attributable to the parameter estimation methods that were used in 1981. In his study, Ree possibly (not reported in the article) used joint maximum likelihood estimation (JMLE), in which item and θ parameters are estimated concurrently, given that it was the only parameter estimation method available at that time. Moreover, it is known that JMLE works best when large number of examinees, such as 1,000, and long tests of as many as 60 items are used in item parameter estimation (Baker & Kim, 2004). Thus, it was typical to have inaccurate estimates in small samples. The findings of the present study also confirmed the findings of Chuah et al. (2006), who found that a pre-calibration sample of 300 examinees was sufficient for accurate estimation of examinee θ in CAST. However, the present results indicated that a pre-calibration sample of 150 could also be feasible if sufficient information existed in the bank for the target examinees' θ levels.

The purpose of the present study was to investigate how θ estimation in 3PLM CAT was affected by calibration sample size conditional on bank size. The results indicated that the θ estimates in CAT were robust against the calibration sample size and the bank size, especially when the item bank had high information that matched the target examinees' θ levels. A sample of 150 examinees might be feasible for calibrating items for use in a CAT item bank, and an item bank of 200 high-quality items that provide high information across the θ continuum could also be useful for many purposes. These results contrast with the findings of prior research that suggested sample sizes of 1,000 or more for accurate item parameter estimates. The difference between the present study and prior research is that the prior research was mainly concerned with the accuracy of item parameter estimation in the 3PLM, whereas the present research focused on the accuracy of the person parameter estimates derived through CAT. Apparently, whatever errors occur in item parameter estimates as the result of small samples and/or small item banks do not have a large influence on the person parameter estimates that result from CAT administration.

The findings of the present study have some implications for future item bank development studies in various disciplines. They can provide an empirical basis for CAT researchers, decision makers, and educational institutions in countries where funding sources are limited and finding large numbers of examinees to calibrate items while developing an item bank for CAT is difficult. The present findings will be especially valuable for reducing the cost and time necessary to develop a CAT item bank.

The findings of the present study are limited to the item banks that were used in the study and the item parameter estimates that were obtained by MMLE in the 3PLM. For this reason, they should only be extended to CAT applications and item banks under similar conditions. Moreover, because the 3PLM

was used to calibrate items in the present study, the findings are limited to this unidimensional dichotomous model of IRT. The findings are also specific to the use of Bayesian θ estimation and should be replicated using maximum likelihood methods to estimate θ.

A natural progression of the present study would be to design research studies to validate how the calibration sample sizes suggested in the present study perform in real CAT applications. Moreover, a similar study on calibration sample size could be conducted using item banks with uniform BIFs that cover a wide range of θ levels. Conducting a similar study using multidimensional models as well as polytomous IRT models would also be useful for future CAT applications.

References

Akour, M., & Al Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, *5*(2), 291–301.

Baker, F. B. (1998). An investigation of the item parameter recovery of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153–169. doi:10.1177/01466216980222005

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Chuah, S. C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241–255. doi:10.1207/s15324818ame1903_5

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351–380. doi:10.1207/s15324818ame1804_2

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, *46*(1), 11–21. doi:10.1177/0013164486461002

Guyer, R., & Thompson, N. A. (2011). *User's manual for Xcalibre 4.1*. St. Paul, MN: Assessment Systems Corporation.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147–200). Washington, DC: American Council of Education.

Hambleton, R. K., & Swaminathan H. (1985). *Item response theory: Principals and applications*. Norwell, MA: Kluwer Nijhoff.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*(3), 279–291. doi:10.1177/014662169101500308

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. doi:10.1177/014662168200600301

IBM Corporation. (2011). *IBM SPSS Statistics for Windows*, *Version 20*.0. Armonk, NY: Author.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989–1020. doi:10.1177/001316446802800401

Patsula, L. N., & Gessaroli M. E. (1995, April). *A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement, 5*(1), 11–19. doi:10.1177/014662168100500102

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 249–260). Minneapolis: University of Minnesota.

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved November 10, 2014 from http://echo.edres.org:8080/scripts/cat/catdemo.htm

Swaminathan, H., & Gifford, J. A. (1979). *Estimation of parameters in the three-parameter latent trait model* (Report No. 90). Amherst, MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27–51. doi:10.1177/0146621602239475

Tang, K. L., Way, W. D., & Carey, P. A. (1993). *The effect of small calibration sample sizes on TEOFL IRT-based equating (TOEFL technical report TR-7)*. Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412. doi:10.1007/BF02293705

Weiss, D. J., & Guyer, R. (2012a). CATSim (version 4.0.6) [Software]. St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J., & Guyer, R. (2012b). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J., & von Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG*. St. Paul, MN: Assessment Systems Corporation.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*(2), 275–291. doi:10.1007/BF02294241

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation.