Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

CAT Field-Test Item Calibration Sample Size: how Large is Large under the Rasch Model?

Wei He[1]

[1] Northwest Evaluation Association

Received: 5 June 2015 Accepted: 5 July 2015 Published: 15 July 2015

Abstract

This study was conducted in an attempt to provide guidelines for practitioners regarding the optimal minimum calibration sample size for pretest item estimation in the computerized adaptive test (CAT) under WINSTEPS when the fixed-person-parameter estimation method is applied to derive pretest item parameter estimates. The field-testing design discussed in this study is a form of seeding design commonly used in the large-scale CAT programs. Under such as seeding design, field-test (FT) items are stored in an FT item pool and a predetermined number of them are randomly chosen from the FT item pool and administered to each individual examinee. This study recommends focusing on the valid cases (VCs) that each item may end up with given a certain calibration sample size, when the FT response data are sparse, and introduces a simple strategy to identify the relationship between VCs and calibration sample size. From a practical viewpoint, when the minimum number of valid cases reaches 250, items parameters are recovered quite well across a wide range of the scale. Implications of the results are also discussed.

Index terms— field-test item calibration, calibration sample size, computerized adaptive test, pretest item calibration, WINSTEPS.

1 Introduction

nlike conventional paper-and-pencil tests (PPT), computerized adaptive tests (CATs) operate on the availability of a large pool of calibrated items **??**Glas, 2010). In order for items to be calibrated, they need to go through a field-testing procedure which aims at assigning test items to examinees so that responses can be available for item parameter estimation **??**Gage, 2009). In CAT, one popular field-testing procedure is to seed field-test (FT) items, also called pretest items, in among the operational items. Often, in a seeding design, FT items are stored in an FT item pool, and a predetermined number of them are randomly chosen from the FT item pool and administered to each individual examinee **??**Buyske, 1998). This seeding approach has several advantages, such as preserving the testing mode, obtaining response data in an efficient manner, and reducing the impact of motivation and representativeness concerns related to administration of pretest items to volunteers (Par shall, 1998).

Author: 121 NW Everett Street, Portland. e-mail: wei.he@nwea.org.

Once responses to FT items are collected, items can be calibrated using an estimation method. Today, a number of software packages do this quite well. Examples are the joint maximum likelihood (JML) method implemented by WINSTEPS (Linacre, 2001) and the marginal maximum likelihood (MML) method using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1999). As a key issue in FT calibration is to make sure FT items are on the same scale as the operational items, a linking/scaling strategy needs to be considered as a part of the scope of the FT item calibration process. In general, any linking/scaling procedures available for PPT can be applied to CAT, and choice of a linking strategy can be predetermined for most CAT testing programs given such factors as FT strategy. Meng and Steinkamp (2009), comparing several pretest item linking designs for

a live CAT program by using both simulated and empirical data, suggested that the fixed-personparameter (FP) estimation method outperforms both Fixed-item-parameter (FI) and Common-item linking with Stocking and Lord Transformation (CI) when pretest item response data are sparse. The FP method investigated by Meng and Steinkamp (2009) and in this study is commonly documented in the literature as Stocking's A method (Stocking, 1988), in which pretest items are estimated by fixing the examinee's final ability estimates. As examinees' final abilities are on the same scale as the operational item parameter estimates, the FT items are automatically on the same scale as the operational items. This approach has been widely applied by programs administering CAT exams under the Rasch model to derive pretest item parameter estimates (Meng & Steinkamp, 2009).

As each individual examinee typically responds only to a subset of FT items in an FT item pool, it is expected that FT item response data will be sparse-a challenge to the accuracy of CAT FT item parameter estimates (Ban et al., 2001).The sparseness rate may vary upon the proportion of the number of pretest items that an individual examinee is administered over the total pretest item pool size-the smaller the proportion, the higher the sparseness rate. What's more, a phenomenon called restricted range of ability (Haynie & Way, 1995;Hsu, Thompson, & Chen, 1998; Stocking, 1990) further complicates FT item calibration because item selection in CAT is customized to the examinee's abilities-high-ability examinees tend to get harder items and vice versa for low-ability examinees. If the examinees used for the calibration sample do not vary enough in ability, item calibration results will be adversely impacted (Stocking, 1990). Fortunately, the seeding design which administers FT items at random regardless of the provisional ability estimates largely alleviates this concern.

One practice that alleviates the effects of sparseness of response data on item parameter estimation accuracy is increasing calibration sample size so that only when an item has been administered to a sufficient number of test-takers are its parameters estimated. However, the literature on CAT does not seem to provide a general guideline about how large a calibration sample size needs to be to be deemed sufficient. In the absence of specific recommendations for CAT, it may be helpful to consult equivalent guidelines for PPT. For example, Wright and Stone (1979) recommended using a sample size of approximately 200 when item parameters are calibrated under the Rasch model. Hambleton, Swamina than, and Rogers (1991) suggested that sample sizes of at least 1,000, 500, and 300 are needed to accurately estimate the item parameters of the three-, two-, and oneparameter item response models respectively. In a situation in which CAT FT item response data are sparse and sparseness rates vary as the result of different factors, such as the one discussed above, more studies are needed. What's more, in light of the fact that the Rasch model is widely used in the large-scale statewide assessments (e.g., The Delaware Comprehensive Assessment System, The Oregon's Assessment of Knowledge & Skills) delivered in the form of CAT, this issue merits a thorough investigation.

For this study, CAT pretest items were randomly selected out of a pretest item pool for administration and calibrated under the Rasch model (Rasch, 1960) by using the WINSTEPS and FP linking method. Specifically, this study endeavored to achieve three goals: 1) introducing a simple strategy to identify the calibration sample size; 2) examining how different calibration sample sizes affect pretest item parameter estimate accuracy; and 3) making recommendations regarding the minimal calibration sample needed to achieve reasonable item parameter estimate accuracy.

2 II.

3 Method and Research Design

A Monte Carlo simulation study was conducted to address the above research questions. selection and content balancing, which involved balancing the content of items administered to match a pre-specified desired percentage of content categories. To control the item exposure rate, one out of a set of items that could provide the most information at the current ability estimate was randomly administered to the examinee. The Bayesian estimation method (Owen, 1973) was used initially, with a prior having a certain mean and standard deviation. The maximum likelihood estimation (MLE) method took over when both correct and incorrect responses were available. To pass the test, examinees needed to answer a minimum of 60 items, with content constraints placed on the set of the items. When 95% of the confidence interval around the candidate's current ability did not encompass the cut score, then the pass/fail decision was returned to the candidate. When the confidence interval included the cut score, candidates continued to take the test with the same content constraints until the current ability estimate was over or below the 95% confidence interval on the cut score or a maximum test length of 250 items was reached.

Field test items, seeded into the operational test, were selected for administration at slots randomly decided regardless of provisional ability estimate and content balancing. Each examinee was administered 15 pretest items, and they were randomly chosen out of 150 pretest items. Responses to field test items were not scored.

4 b) Item Pool Characteristics i. Scoreableitem pool

The scoreable item pool used in this study was simulated by mimicking the distribution of a real item pool used by a large-scale computerized adaptive test. The simulated item pool contained 1602 Rasch items distributed in eight content strands with a mean of -0.266 and a standard deviation of 1.76. By "scoreable", it means the responses to these items were counted toward the final ability estimates. Table 1 and Figure **??** present the descriptive statistics and distribution of item difficulties of this scoreable item pool.

The FT item pool consisted of 150 items randomly selected from the scoreable item pool described above. As mentioned previously, the response data for the FT items was sparse because only a subset of items was selected out of the FT item pool. Although randomly assigning FT items to examinees could theoretically ensure that FT items-regardless of their difficulty levels-get a similar level of exposure, it was observed that some items were exposed considerably more than others. Thus, the calibration sample size used in this study was decided by the minimum number of valid cases (denoted as VCs hereafter) that each field test item needed to contain.

To identify how different calibration sample sizes yielded different VCs, a simulation study was conducted first, in which pretest item selection procedure (i.e., random selection) was mimicked using the pretest item pool only. Specifically, the predetermined number of FT items was administered to target examinee populations of different sizes, and then the number of VCs that each pretest item contained was counted given a specific calibration sample. The simulation results revealed that, to make sure that each field test item contained at least 1000, 500, 250, 120, 60, or 30 responses respectively, the calibration sample sizes had to reach 11000, 6000, 3000, 1500, 850, or 470 correspondingly. In other words, given that 15 items were selected out of a 150-item FT pool, the approximate ratio between calibration sample size and VC was between 10 and 12. Table 3 indicates the relationship between calibration sample size and VCs for each FT item. 1. The pre-specified number (denoted as N) of examinees under "Calibration Sample Size" in Table 3 was randomly drawn out of the distribution with the mean of -.029 and the standard deviationof.4852. This distribution mimicked the target examinees' ability distribution for a largescale CAT program. Each examinee was administered 15 FT items randomly drawn out of the FT item pool. This step yielded a sparse person-byitem response dataset of size N*150. 2. The computerized adaptive testing algorithm described under the CAT Model section was run to get an estimated ability for each of N examinees. This step yielded N ability estimates. 3. WINSTEPS was used to calibrate the FT items under default settings by fixing the estimated abilities obtained in 2). 4. Steps 1), 2), and 3) were replicated 100 times, resulting in 100 sets of item parameter estimates.

5 e) Analysis

The analysis for each field test item was focused on its calibration accuracy and precision, measured by bias, absolute bias (Abias), and mean squared error (MSE). Following are the equations used to compute the above statistics. Let k=1,2,?,100 replications and j= 1,2, ?,100 items. and denote the item difficulty parameter, i.e., true item difficulty parameter and item difficulty parameter estimate respectively:? Bias 100 /) ( ) (100 1 ? ? ? ? ? ? ? = ? = k j kj j b b Bias Eq[1] ? Abias 100 / |) (|) (100 1 ? ? ? ? ? ? ? ? = ? = k j kj j b b Abias Eq [2] ? MSE 100 /) ( ) (100 1 2 ? ? ? ? ? ? ? ? = ? = k j kj j b b MSE Eq [3] III.

6 Results

The FP method is criticized for introducing errors in calibrating the FT items because it treats ability estimates as true abilities to maintain the scales of subsequent item pools (Ban et al., 2001), but estimated abilities may be different from true abilities. To ensure this is not a concern in the current study, true and estimated abilities were reported in Table 4. What's more, average bias, average MSE, and correlation coefficient between estimated and true abilities ( ) were also computed and presented in Table 5. These statistics indicate that examinees' abilities were recovered very well with almost unbiased average ability estimates and low estimation errors. The average test length was 107 items. The procedures used for field test item calibration were described as follows. For each calibration sample size, the calibration procedure remained the same. One hundred replications were run for each calibration sample size.

Year 2015 For some items, when the calibration sample size was small, there were some runs failing to yield valid item parameter estimates due to perfect scores, i.e., all of the responses to a certain item are either correct or incorrect. In the case of perfect scores, WINSTEPS can still report the item parameter estimates, but with very substantial standard errors. Thus, this study did not count a run as valid if the run involved estimating perfect scores.

Figure 3 demonstrates the relationship between the number of runs yielding no available item parameter estimates and the item difficulty parameter. Clearly, the situation in which item parameter estimates were unavailable was more likely to occur with those items at the tails of the scale, in particular, easy items. Increasing the calibration sample size seemed to minimize the occurrence of the above situation. For example, when the calibration sample size was 470, item parameter estimates failed to be reported for 43 items in certain runs. However, only 6 items encountered the same problem when the calibration sample size was 1500. Bias. The magnitudes of the bias produced by different calibration sample sizes are plotted against true item difficulty parameter in Figure **??**. In general, these plots indicate that easy items tend to be underestimated and vice versa for hard items. With the increase of VCs for each item, we can see that the magnitude of the bias became less pronounced. From the practical viewpoint, when a calibration sample size allowed VCs to reach 250, the bias for item parameter estimates was negligible for items with log its between -3 and 3. When a calibration sample allowed VCs to reach 1,000, item parameter estimates were almost unbiased. Table 6 also provides summary statistics about the absolute bias of item parameter estimates given by different VCs. Clearly, absolute bias also decreased with the increase of calibration sample size. Wright and Douglas (1977) proposed a simple bias correction method that can be used to remove the bias in an item parameter estimate using the JML method.

In WINSTEPS, this method is implemented by a command called STBIAS, which involves multiplying the item parameter estimate by the correction factor (L-1)/L, where L is the test length. By default, STBIAS is not invoked in WINSTEPS unless it is set as Y. Wang and Chen (2005) reported that STBIAS can significantly reduce the magnitudes of the bias in item parameter estimation. To examine how the magnitude of bias was slightly improve item parameter estimates by yielding a slightly lower average absolute bias and reducing the spread of item parameter estimates. Figure 5 compares the average bias for item parameter estimates when STBIAS is and is not used. corrected by STBIAS for sparse response data like that in this study, item estimation was conducted by implementing STBIAS, and the magnitude of the bias in the item parameter estimate when STBIAS was not used was compared with that when STBIAS was used. The results, illustrated in MSEs for item parameter estimates exhibited very similar patterns to those for bias. Specifically, both easy and hard items tend to be associated with larger errors than items in the middle of the scale, particularly when calibration sample size yielded VCs lower than 250. When VC reached 250 and beyond, it is clear that the magnitudes of MSEs were negligible even for items with difficulty value beyond 3 log it in absolute value. Figure 6

7 Discussion and Conclusions

As mentioned previously, pretest item response data tend to be sparse under a seeding design in which only a subset of items is selected for administration in the CAT. Additionally, as FT items are likely to be exposed at different rates-some items receive more administrations than others, the question arises as to how large the calibration sample size needs to be so that item parameters are estimated accurately. This study was conducted in an attempt to provide practitioners certain guidelines about the optimal minimum calibration sample size for CAT pretest item estimation under WINSTEPS when the fixed-personparameter estimation method is applied to derive pretest item parameter estimates.

Under such a design, as demonstrated, different calibration sample sizes lead to different average VCs given the ratio being fixed between the number of FT items administered to each examinee and the total FT item pool size. As expected, the larger the calibration sample size is, the larger the numbers of VCs are, and thus the better items are calibrated. This study recommends that, when the FT response data are sparse, focus should be placed on the valid cases that each item may end up with given a certain calibration sample size. As the methodology introduced in this study indicates, the relationship between VCs and calibration sample size can be very easily identified simply by simulating the operational FT item selection procedure using the FT item pool only. From a practical viewpoint, when the minimum number of valid cases reaches 250, item parameters are recovered quite well across a wide range of the scale. This number seems to be in agreement with, though slightly higher than, what Wright and Stone (1979) recommended-a sample size of approximately 200 for a paper-and-pencil test.

Clearly, the ratio between the number of FT items administered to each examinee and the total FT item pool size plays a key role in deciding the calibration sample size. The smaller the ratio is, the smaller the calibration sample size is needed. Collecting responses from a large sample may not be an issue for largevolume testing programs, but may be so for smallvolume ones. Thus, to help item throughput, it is recommended to keep this ratio to a low number given the use of the same field-testing and calibration procedure.

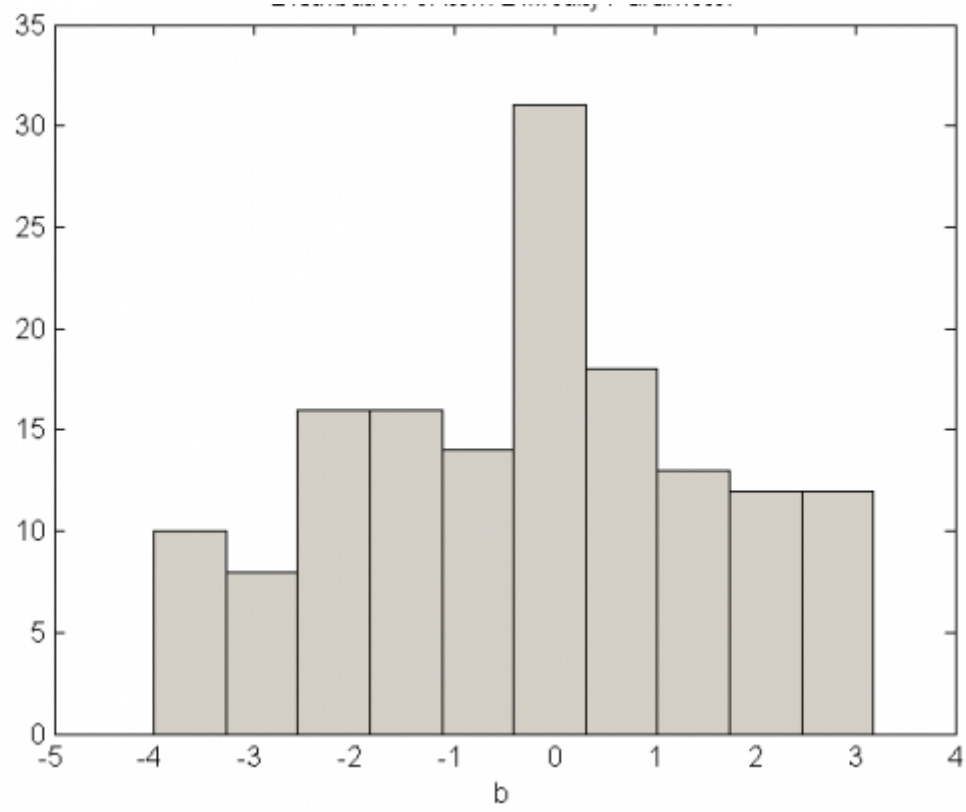
Unlike what is reported in Wang and Chen (2006) in which biases of item parameter estimates are significantly corrected by the STBIAS command especially in the extreme situations, the STBIAS command only slightly improved estimate accuracy in the current study. A close look at the results revealed that L was defined as 150 (i.e., the total number of the items in the item pool) rather than the actual number of items (i.e., 15 items) administered to each examinee when STBIAS was set as Y. Clearly, if L is a large number, (L-1)/L tends to approach unity, thus playing a weaker role in bias correction. Therefore, given the situation in which a large calibration sample is unaffordable and STBIAS is in need to improve item estimate accuracy, it is not recommended to administer items out of a large FT item pool. This recommendation is tied up with keeping a reasonable ratio as discussed above.

As mentioned in the Results section, the FP method has the potential to introduce errors in calibrating the FT items especially when ability estimates are inaccurate. The CAT model mimicked in this study is a pass/fail classification test, implying that ability estimates near the cut score may be fairly inaccurate and thus provide a poor linking. This does not seem to be a concern in this study, as Table 5 indicates that ability estimates are recovered quite well. The fact that the average test length (i.e., 107 items) is considerably long plays a key role. However, it is anticipated that poor ability estimates may produce a poor linking, thus challenging the results in this study. Future research should be conducted along this line to examine how ability estimates affect item parameter estimate accuracy in such a seeding FT item design in the CAT.

Investigation into item parameter estimation accuracy was conducted in this study by considering calibration sample size as the only affecting factor. In reality, such factors as FT item position or calibration sample distribution also exert impacts. Future research should look at how these factors interact with each other to affect estimate accuracy. Additionally, item calibration was conducted by using only one linking design and estimation method. Adding different linking designs and estimation methods, in conjunction with the factors mentioned above, also merits further research.

2

Figure 1: Figure 2 :



5

Figure 2: - 5 .

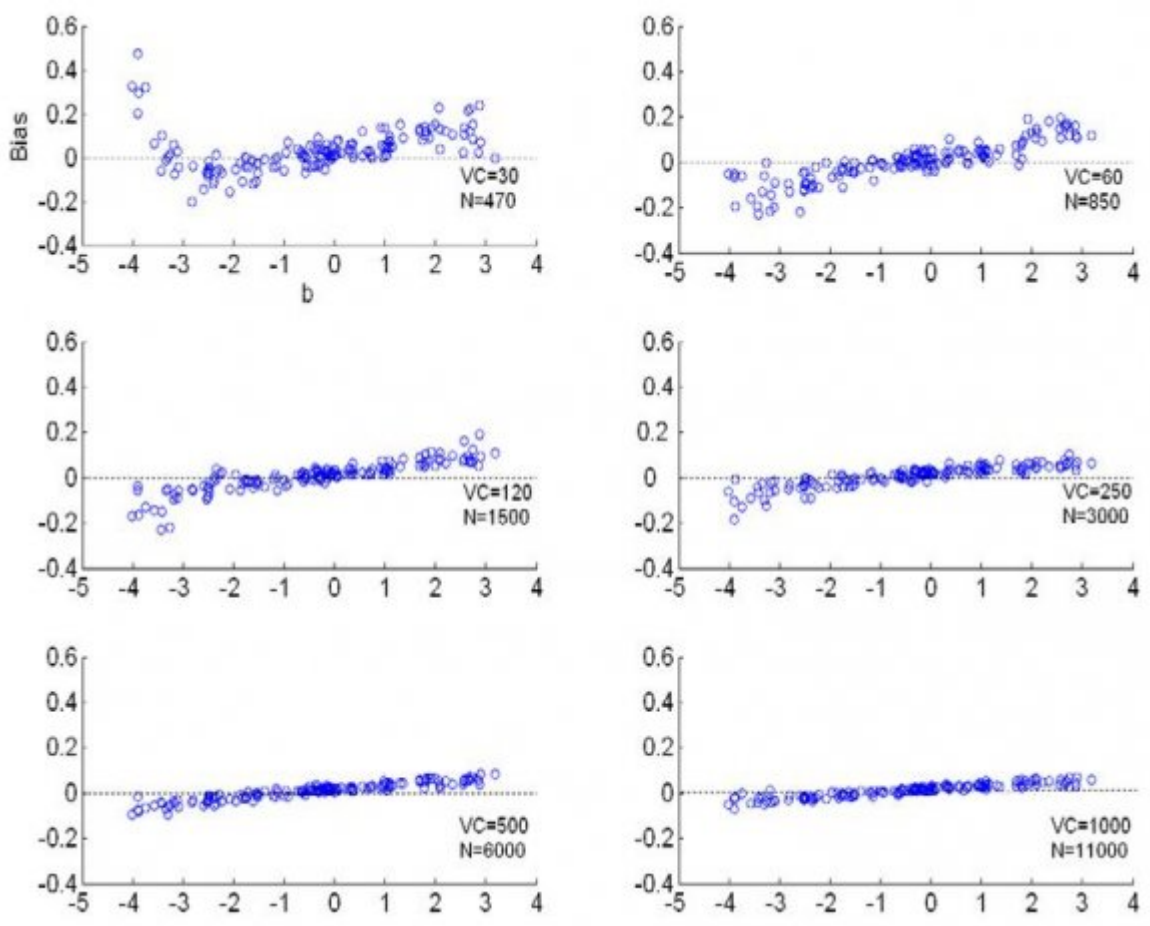
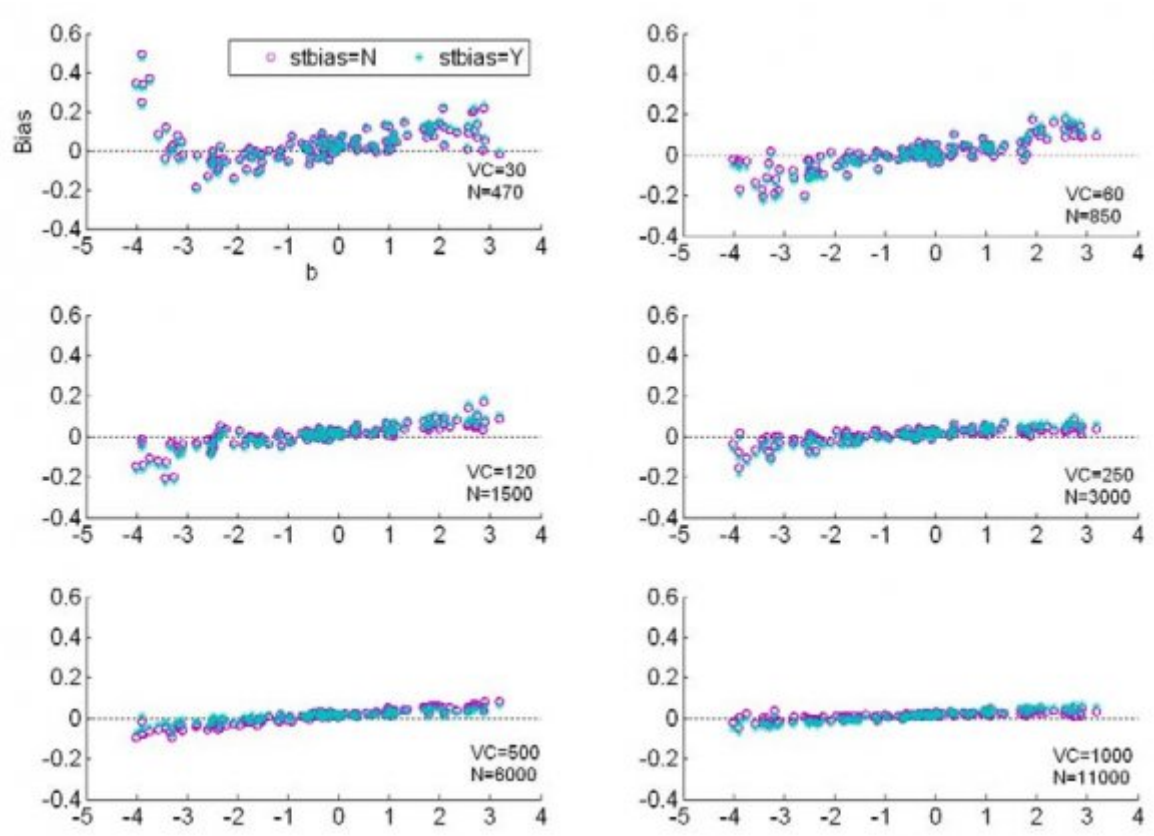
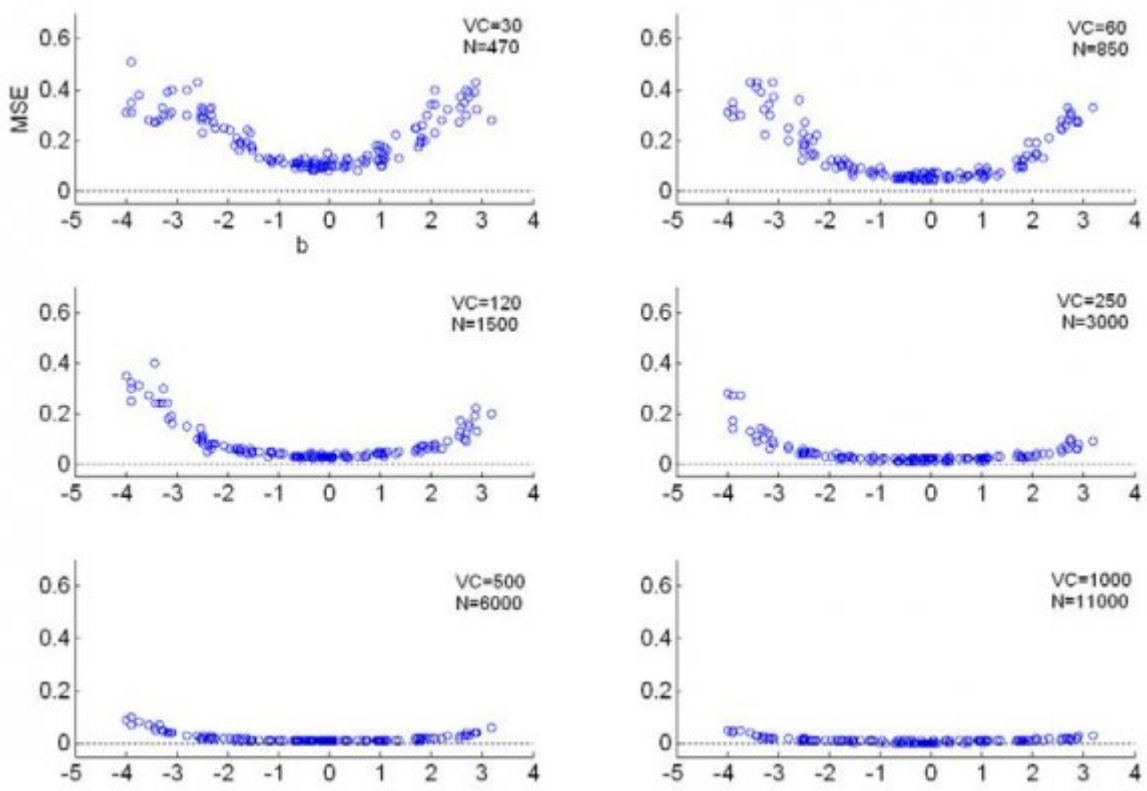

Bias
b
VC=30
N=470
VC=60
N=850
VC=120
N=1500
VC=250
N=3000
VC=500
N=6000
VC=1000
N=11000

Figure 3:



3

Figure 4: Figure 3 :



20154

Figure 5: Year 2015 Figure 4 :

2

and Figure 2 present the
descriptive statistics and distribution of item difficulties
of this FT item pool. These FT items spanned a wide
range of the ability scale.

Figure 6: Table 2

1

	Total Number	Mean	Std. Deviation	Minimum	Maximum
b	1602	-0.266	1.760	-4.418	3.301

[Note: Figure 1 : Scoreable item difficulty distribution]

Figure 7: Table 1 :

2

	Total Number	Mean	Std. Deviation	Minimum	Maximum
b	150	-0.340	1.817	-4	3.19

Figure 8: Table 2 :

3

Calibration Sample Size	11000	6000	3000	1500	850	470
VC	1000	500	250	120	60	30

Figure 9: Table 3 :

4

	Mean	Std. Deviation	Maximum	Minimum
?	-0.003	0.505	1.528	0.010
? ?	0.021	0.568	1.836	-1.853

Figure 10: Table 4 :

5

Figure 11: Table 5 :

6

VC/Calibration sample	Maximum	Minimum	Mean	Std. Deviation
30/470	.472	.000	.069	.071
60/850	.196	.001	.062	.057
120/1500	.192	.000	.047	.044
250/3000	.100	.000	.035	.028
500/6000	.083	.000	.031	.022
1000/11000	.069	.001	.026	.017

Figure 12: Table 6 :

7

Year 2015

Figure 13: Table 7 ,

7

		Estimates						
VC	Mean STBIAS=N STBIAS=Y	STBIAS=N	STBIAS=Y	STBIAS=N	STBIAS=Y	STBIAS=N	STBIAS=Y	STBIAS=N STBIAS
30	0.069	0.065	0.071	0.074	0.493	0.493	0.000	0.000
60	0.062	0.053	0.057	0.050	0.177	0.205	0.000	0.000
120	0.047	0.039	0.044	0.037	0.172	0.205	0.000	0.000
250	0.035	0.028	0.028	0.022	0.081	0.156	0.000	0.000
500	0.031	0.023	0.022	0.015	0.062	0.075	0.000	0.000
1000	0.026	0.019	0.017	0.012	0.051	0.051	0.000	0.000

Note. N represents calibration sample size

Figure 14: Table 7 :

[Applied Psychological Measurement] , *Applied Psychological Measurement* 12 (3) p. .

[Hsu et al. ()] , Y Hsu , T D Thompson , W.-H Chen . 1998.

[Linacre ()] , J M Linacre . 2001. (Rasch measurement computer program (Version 3.31) [Computer software]. Chicago: Winsteps.com)

[Ban et al. ()] 'A comparative study of on-line pretest item: Calibratoin/Scaling methods in computerized adaptive testing'. J-C Ban , B A Hanson , T Wang , Q Yi , D J Harris . *Journal of Educational Measurement* 2001. 38 (3) p. .

[Meng and Steinkamp ()] 'A comparison study of CAT pretest item linking designs'. H Meng , S Steinkamp . *Paper presented at the 74th annual meeting of the psychometric society*, 2009.

[Kingsbury ()] 'Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test'. G G Kingsbury . *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, D J Weiss (ed.) (the 2009 GMAC Conference on Computerized Adaptive Testing) 2009. (Retrieved from www.psych.umn.edu/psylabs/CATCentral/)

[Haynie and Way ()] *An investigation of item calibration procedures for a computerized licensure examination. Paper presented at symposium entitled Computerized Adaptive Testing at the annual meeting of NCME*, K A Haynie , W D Way . 1995. San Fancisco.

[Wright and Douglas ()] 'Best procedures for sample-free item analysis'. B D Wright , G A Douglas . *Applied Psychological Measurement* 1977. 1 p. .

[Wright and Stone ()] 'Best test design'. B D Wright , M H Stone . *Chicago: Measurement, Evaluation* 1979. Statistics, and Assessment Press.

[Zimowski et al. ()] *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items*, M F Zimowski , E Muraki , R J Mislevy , R D Bock . 1999. (Computer program]. Chicago: Scientific Software International)

[Van Den Wollenberg et al. ()] *Consistency of Rasch model parameter estimation: a simulation study*, A L Van Den Wollenberg , F W Wierda , P G W Jansen . 1988.

[Jansen et al. ()] 'Correcting unconditional parameter'. P G Jansen , A L Van Den Wollenberg , F W Wierda . *Applied Psychological Measurement* 1988. 12 (3) p. .

[Hambleton et al. ()] *Fundamentals of Item Response Theory*, R K Hambleton , H Swamina Than , H J Rogers . 1991. Newbury Park, CA: Sage.

[Par Shall ()] *Item development and pretesting in a computer-based testing environment. Paper presented at the colloquium Computer-Based Testing: Building the Foundation for Future Assessments*, C G Par Shall . 1998. Philadelphia, PA.

[Wang and Chen ()] 'Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models'. W C Wang , C T Chen . *Educational and Psychological Measurement* 2005. 65 (3) p. .

[Paper presented at the Annual Meeting of the National Council on Measurement in Education] *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, San Diego.

[Glas (2003)] *Quality control of online calibration in computerized assessment. Law School Admission Council Computerized Testing Reports 97-15*, C A W Glas . `http://www.lsac.org/lsacresources/Research/CT/CT-97-15.pdf` 2003. September 2003.

[Stocking ()] *Scale drift in on-line calibration*, M L Stocking . 1988. Princeton, NJ: ETS. (Research Rep. 88-28)

[Stocking ()] 'Specifying optimum examinees for item parameter estimation in item response theory'. M L Stocking . *Psychometrika* 1990. 55 (3) p. .