Dealing with Variability within Item Clones in Computerized Adaptive Testing

Research Report

Chingwei David Shin
Yuehmei Chien

May 2013

ALWAYS LEARNING PEARSON

About Pearson
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education. For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research Reports
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world. Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders.

Abstract

Item cloning is a technique that generates, on the fly, a large set of operational items from an item template. Item cloning greatly reduces the cost of item development, increases the number of operational items available, and enhances item security. However, some researches have shown that item clones from the same parent might not be completely psychometrically isomorphic, therefore causing variability in item parameters of clones.

The purpose of this study is to examine different means of dealing with the variability of clone items in order to moderate the possible loss of precision in ability estimation in CAT. To possibly capture the difference in item parameters among clones, two functions are investigated. First, the expected response function (ERF), which models the random variations of item clones in item parameters within the family, is considered; with the ERF, a close-fitting three-parameter logistic (3PL) item-characteristic curve can be obtained and used in CAT. Second, the expected item information functions (EIIF) are proposed in this study to obtain a close-fitting 3PL item information function. In addition to the ERF and EIIF, three different item selections are also evaluated in this study: the maximum item information (MII) method; the minimum expected posterior variance (MEPV) method, and the maximum posterior weighted information (MPWI) method. To examine the effect of item cloning on measurement precision in CAT, a simulation study was designed and conducted.

The main findings of this study are: first, when the items are sufficient for certain ability levels, the loss of precision is not observed or is small; and second, the ERF parameters are promising in terms of reducing the possible precision loss due to the lack of isomorphic clones and the lack of adaptive item templates in the pool.

Keywords: item cloning, item templates, CAT

Dealing with Variability within Item Clones in Computerized Adaptive Testing

Introduction

Item cloning is a technique that generates, on the fly, a large set of operational items from an item template, in which the template is called a parent template and those operational items generated by a parent template are called (item) clones. Item cloning greatly reduces the cost of item development, increases the number of operational items available, and enhances item security. However, some researches (Enright, Morley, and Sheehan, 2002; Meisner, Luecht, and Reckase, 1993) have shown that item clones from the same parent might not be completely psychometrically isomorphic, therefore causing variability in item parameters of clones.

It is conceivable that, if the family structure is ignored, the variability within item clones might erode the item-calibration precision to some degree and might also reduce measurement precision to the test for which those item parameters are used. However, the research results by Glas and van der Linden (2003) showed that the effect of item cloning on item calibration was not pronounced; but on computerized adaptive testing (CAT), small precision erosions were observed. The measurement precision, eroded at some level by item cloning in CAT, was confirmed in another study by Bejar, Lawless, Morley, Wagner, Bennett, and Revuelta (2003). Therefore, in order to use item cloning in CAT, it is important to find a way to incorporate the different levels of variability in CAT to avoid the possible precision loss.

The purpose of this study is to examine different means of dealing with the variability of clone items in order to moderate the possible loss of precision in ability estimation in CAT. To possibly capture the difference in item parameters among clones, two functions are investigated. First, the expected response function (ERF) (Lewis, 1985; Mislevy, Wingersky, and Sheehan, 1994), which models the random variations of item clones in item parameters within the family,

is considered; with the ERF, a close-fitting three-parameter logistic (3PL) item-characteristic curve can be obtained and used in CAT. Second, the expected item information functions (EIIF) are proposed in this study to obtain a close-fitting 3PL item information function. (More detailed information is provided in a later section.)

In CAT, how the item is adaptively selected is critical. Therefore, three different item selections are also evaluated in this study: the maximum item information (MII) method; the minimum expected posterior variance (MEPV) method, in which the expectation is taken over the posterior predictive distribution of the responses to a random item from the family (Glas and van der Linden, 2003); and the maximum posterior weighted information (MPWI) method, in which the integral is taken over a finite interval of ability values (Veerkamp and Berger, 1997; van der Linden and Pashley, 2010).

There are two research questions that are intended to be answered through this study. First, is ERF or EIIF a good solution to model the variability within family in CAT? Second, do the different item selection methods—the MII, the MEPV, and the MPWI—make differences on measurement precision? To answer these two research questions, a simulation study was designed and conducted.

The Model

In item response theory (IRT), the item response function (IRF; the realization of IRF is called item characteristic curve, or ICC) defines the probability that a person with a specific ability level can answer a specific item correctly. After the item is statistically calibrated, the item parameters that determine the shape of the IRF can be obtained. Ideally, the item parameters should be accurate in order to be used in a real test situation. However, the ideal situation might not be attainable due to various reasons, such as the small sample size and the lack of fit between

the model and the data. The ERF is one of the approaches to deal with the imperfectly known item response functions, which take into account uncertainty about item parameters. Charles Lewis first introduced the concept of ERF in 1985 (and he formally gave a description of ERF in 2001). After ERF was introduced, it was applied to various contexts in measurement (Mislevy, Sheehan, and Wingersky, 1993; Mislevy, Wingersky, and Sheehan, 1994; Bejar, Lawless, Morley, Wagner, Bennett, and Revuelta, 2003). The concept of ERF perfectly fits to the context of item cloning since the randomness in the process of generating an item clone introduces the uncertainty about the "true" item parameters for the clone. Therefore, indeed, item cloning can be modeled by the ERF.

The ERF

The ERFs are actually the average probabilities over all generated clones within the family. The closed-form solution of ERF (which is the way to obtain the exact item parameters of the expected probability) does not exist in IRT. However, the expected probabilities can actually be fitted to the closest IRF, which is called fitted ERF (Mislevy, etc., 1994), and therefore the item parameters calibrated by ERF are obtained. Figure 1 shows the ICCs of 50 sampled clones in gray, the ICC of the fitted ERF based on those 50 clones in green, and the expected IRF in red circles. As shown in Figure 1, the fitted ERF is flatter than those IRFs of clones, which means the fitted ERF has smaller discrimination power then those clones.

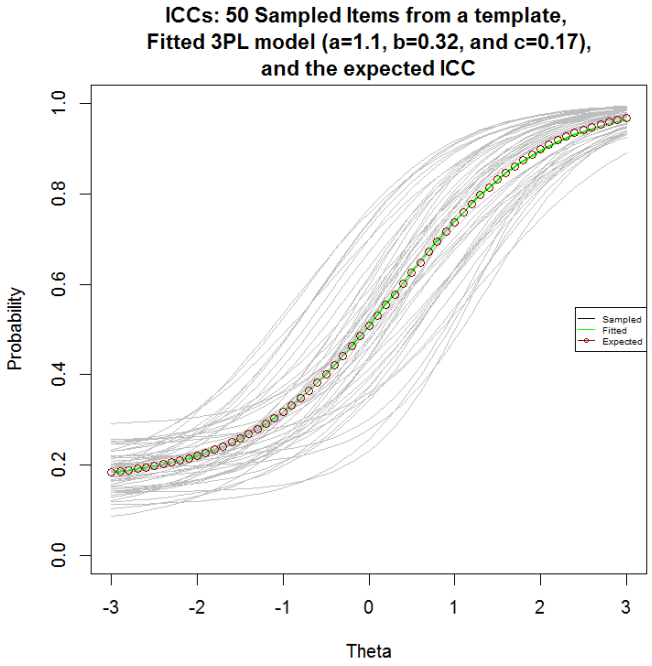

Figure 1. The fitted ERF based on 50 clones in green and the expected IRF.

The EIIF

As mentioned previously, in CAT, it is important to adaptively select the next item that best suits the test taker's ability based on her/his current ability estimate. The amount of the Item information is mostly used as the item selection criteria in CAT. Therefore, in addition to the ERF, the expected item information functions (EIIFs) are proposed to replace the regular item information in CAT. Similarly to the ERF, a close-fitting 3PL item information curve can be obtained using the EIIF.

Method

Simulation Conditions

The instances of items generated by the same parent template are clones in the same family. In practice, calibrating each clone parameter is an arduous effort and is actually impossible for those families that have an unlimited number of clones. However, through

simulation the clone parameters are available to be generated. In addition to the generated clone parameters, three other different kinds of item parameters were generated and used in this study. Those were labeled and described below:

1. PARENT—the template parameters that ignore the variability within templates and use the data from the family to calibrate and obtain a single set of discrimination, difficulty, and guessing parameters. (In this study, PARENT parameters are taken from a real pool). This set of item parameters is the worst case since the variability within templates is completely ignored.
2. ERF—the template parameters obtained by using the ERF to find the close-fitting 3PL item parameters.
3. EIIF—the template parameters obtained by using the EIIF to find the close-fitting 3PL item parameters.
4. CLONE—the individual clone parameters (each clone has its own set of discrimination, difficulty, and guessing parameters). This set of parameters is deemed to be the "true" item parameter set which is not observable in reality but available in the simulation situation.

In total, there are twelve study conditions with four different kinds of item parameters and three item selection methods.

Data and Simulation Design

One real item pool from a large-scale CAT program served as PARENT parameters, in which items were calibrated using the regular 3PL model. This pool contains 309 items in which mean discrimination =1.0 and standard deviation (SD) = .35, mean difficulty =.21 and SD = 0.98, and mean guessing =.15 and SD = 0.08. There are some items in the pool that should not

appear in the same test because item stems are overlapped or one item might give clues to another. These kinds of items are put into the same group named "overlap group," in which a test cannot include more than one item from the same overlap group. This pool contains thirty-nine overlap groups with two to eleven items in each group.

To generate the CLONE parameters from the parent, the mean of item parameters (discrimination, difficulty, and guessing) and the variance-covariance matrix of item parameters have to be known for each family. The mean vector can be taken from its PARENT parameters. The variance-covariance matrix contains the information about the magnitude of variability within each family. The magnitude of variability information used in this study was referred to in a companion study conducted by Lathrop and Chang (2013) which analyzed the magnitude of variability within templates using real data from an online template-driven assessment. In the study of Chang et al., it was found that the magnitude of variability was varied across templates. Some templates showed more variability in the difficulties than others. Based on this finding, this study randomly and equally assigned large, middle, or small variability to parent templates and the resulting proportion for each of the three different levels of variability was 1/3. The study of Chang et al. also found that the large variability in terms of standard deviation of difficulty was about .6 or .7, the middle variability was about .3 or .4, and the small variability was about .1 or .2. Therefore, the variances of item difficulty in the variance-covariance matrices were assumed to be .45 (which is about the average of square of .6 and .7), .13, and .025 based on the standard deviation of difficulty within family from the real data analysis. The covariance structures with three different levels of variability are:

$$\sum_S = \begin{bmatrix} .01200 & .00111 & .00027 \\ .00111 & .02500 & .00112 \\ .00027 & .00112 & .00033 \end{bmatrix}$$

$$\sum_M = \begin{bmatrix} .01500 & .00478 & .00068 \\ .00478 & .13000 & .00280 \\ .00068 & .00280 & .00082 \end{bmatrix}$$

$$\sum_L = \begin{bmatrix} .02000 & .00478 & .00068 \\ .00478 & .45000 & .00280 \\ .00068 & .00280 & .00082 \end{bmatrix}$$

Two hundred sets of CLONE parameters were drawn for each family from the multivariate normal distribution with a mean vector from their corresponding PARENT parameters and with one of the covariance structures as listed above. Based on those two hundred sets of CLONE parameters for each family, the ERF close-fitting 3PL parameters and the EIIF close-fitting 3PL parameters were also obtained for each family.

For each study condition, two kinds of samples were generated to obtain the overall results and the results conditioning on different theta levels. First, a normal sample of 5,000 simulees was generated from a standard normal distribution. Second, the conditional sample of 7,000 simulees was also generated, with 1,000 at 7 equally spaced theta levels from -3.0 to 3.0. For the first item selection, the initial ability is set to -1.0. The test length is 20.

Results

Table 1 and Figures 2 through 4 present the bias for the normal samples and the conditional samples for each of the studied conditions. The CLONE parameters generally had less bias than other three types of item parameters across three different item selection methods with two exceptions at theta = 0 and at theta = 1. Normally speaking, the CLONE parameters should always have less bias because they are the “real” parameters for the cloned items. However, the results show that when theta = 0 and theta = 1, the EIIF and/or ERF have less bias than the CLONE parameters. To explain this finding, the pool structure of the PARENT parameters, which are the means of the CLONE parameters, was examined and shown in Figure

5. Obviously, there were more parent templates with difficulty values around 0 or 1, and, among those parent templates, some had a relatively higher discrimination than other templates in other difficulty ranges. This finding indicates that when there are sufficient items or templates adaptive to the test taker's ability, the bias in the adaptive tests using the EIIF or the ERF parameters was not greater than those using the CLONE parameters.

To compare the ERF, EIIF, and PARENT parameters, which are practically available, the smallest bias value in Table 1 for each of the different true thetas or for the true thetas generated from a standard normal distribution are shown in italics and highlighted with a light background color. It is clearly shown that among the three different kinds of item parameters, the ERF parameters generally performed best in terms of smaller bias and the EIIF parameters performed the second. By observing Figures 2 to 4, it is found that the three different kinds of parameters performed similarly for those thetas smaller than zero, but the PARENT parameters had a larger bias than the ERF and the EIIF parameters for thetas equal to or greater than zero. By examining Table 1, it is found that the ERF parameters consistently produced less bias for the normal true theta distribution across three different item selection methods and the PARENT parameters consistently produced a larger bias for the normal true theta distribution and the conditional thetas.

Table 1

Bias in Ability Estimates

Item Selection Method	Case of Item Parameters	Theta -3	-2	-1	0	1	2	3	Standard Normal
MII	CLONE	-0.4273	-0.1302	-0.0296	0.0155	0.0240	0.1075	0.3062	0.0015
	PARENT	-0.5533	-0.2672	-0.0811	0.0402	0.0748	0.3471	0.5158	0.0154
	ERF	*-0.5436*	*-0.2589*	*-0.0622*	0.0184	0.0192	*0.2210*	0.4771	*0.0037*
	EIIF	-0.5761	-0.3055	-0.1108	*-0.0056*	*0.0016*	0.2552	*0.4416*	-0.0327
MPWI	CLONE	-0.4534	-0.1422	-0.0220	0.0103	0.0156	0.1036	0.3425	-0.0019
	PARENT	-0.6061	-0.2895	-0.0830	0.0476	0.0796	0.3295	0.5158	0.0145
	ERF	*-0.5619*	*-0.2795*	*-0.0536*	0.0120	0.0069	*0.2128*	0.4786	*0.0001*
	EIIF	-0.5999	-0.3172	-0.1134	*-0.0018*	*-0.0058*	0.2463	*0.4583*	-0.0365
MEPV	CLONE	-0.4094	-0.1294	-0.0308	0.0173	0.0219	0.1150	0.2995	0.0015
	PARENT	-0.5438	-0.2685	-0.0745	0.0451	0.0949	0.3451	0.4970	0.0255
	ERF	*-0.5348*	*-0.2553*	*-0.0709*	0.0201	0.0237	*0.2122*	0.4709	*-0.0041*
	EIIF	-0.5761	-0.3120	-0.1199	*-0.0030*	*0.0101*	0.2559	*0.4188*	-0.0298

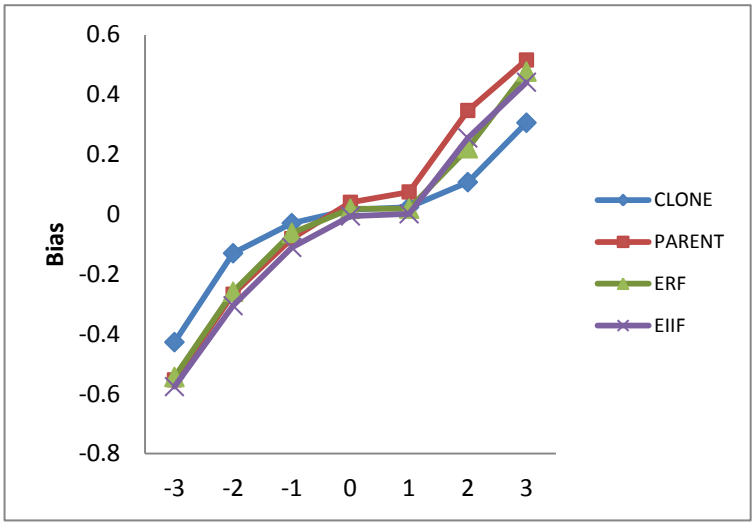

Figure 2. The bias results of the MII method.

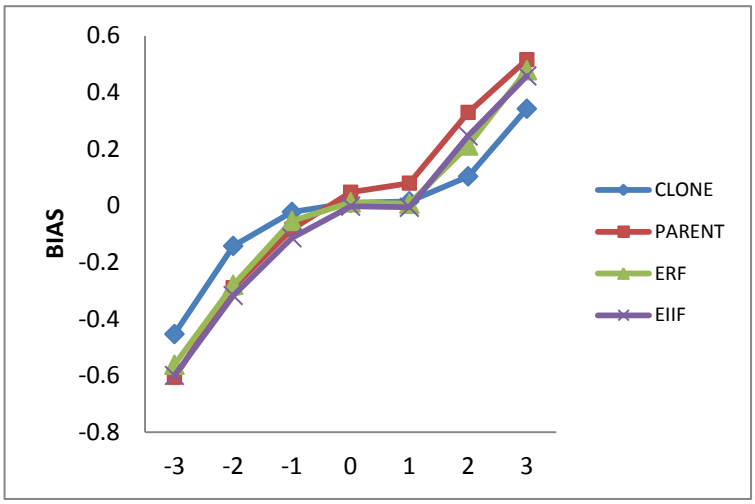

Figure 3. The bias results of the MPWI method.

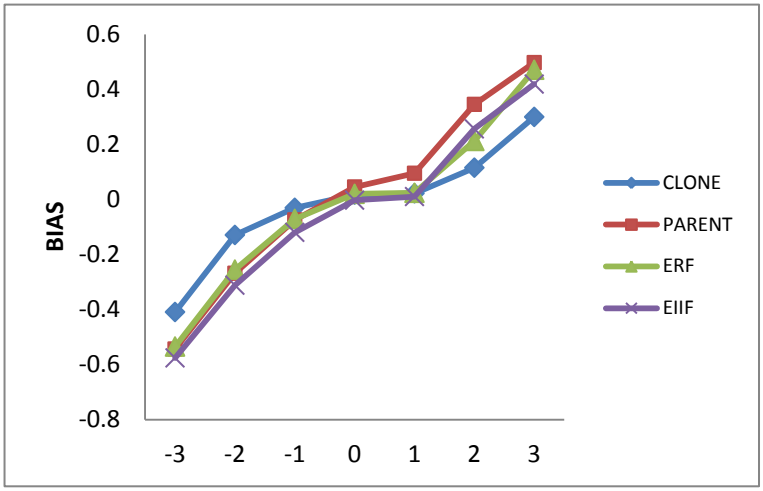

Figure 4. The bias results of the MEPV method.

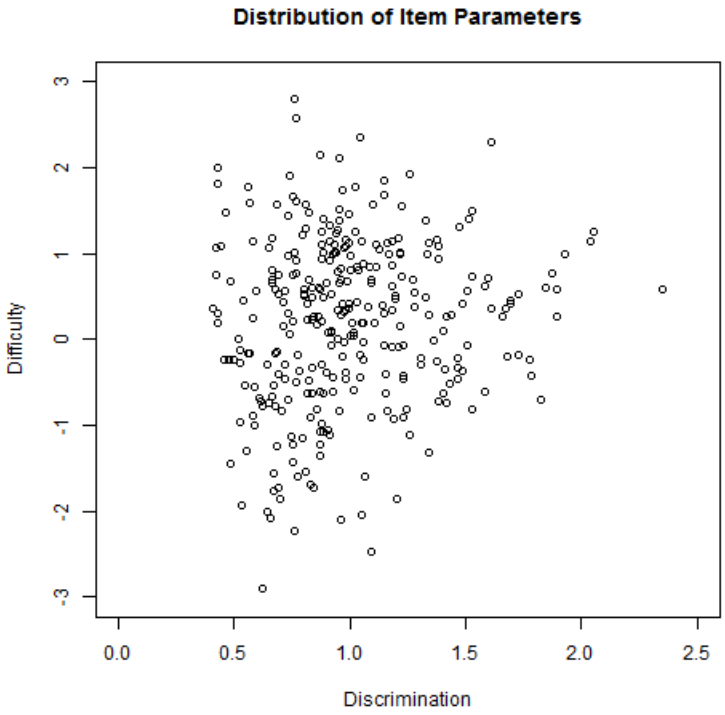

Figure 5. The distribution of item difficulty and discrimination of the pool.

The mean square error in ability estimates are shown in Table 2 and Figures 6 to 8. The CLONE parameters generally had smaller MSE values than the other three types of item parameters across three different item-selection methods with one exception at theta = 1. This finding (similar to what was found in the bias results) indicated that when there are sufficient items or templates adaptive to the test taker's ability, the measurement precision in the adaptive tests using the ERF parameters or the EIIF parameters or even the PARENT parameters were not eroded when comparing with those using the CLONE parameters.

To compare the ERF, EIIF, and PARENT parameters, which are practically available, the smallest MSE values in Table 2 for each conditional theta or for the normal true thetas distribution are shown in italic and highlighted with a light background color. Generally, among

the three different kinds of item parameters, no one consistently performed better than the others across different item-selection methods and across different theta levels.

Table 2

Mean Square Error in Ability Estimates

Item Selection Method	Case of Parameters	Theta							
		-3	-2	-1	0	1	2	3	Standard Normal
MII	CLONE	0.2879	0.0908	0.0552	0.0384	0.0416	0.0753	0.1598	0.0508
	PARENT	0.4199	0.1504	0.0706	*0.0440*	0.0387	0.1921	0.3456	0.0661
	ERF	*0.4138*	*0.1441*	*0.0579*	0.0560	0.0355	*0.1227*	0.3086	*0.0653*
	EIIF	0.4583	0.1803	0.0652	0.0547	*0.0363*	0.1375	*0.2748*	0.0658
MPWI	CLONE	0.3207	0.0950	0.0541	0.0366	0.0454	0.0774	0.1873	0.0504
	PARENT	0.4940	0.1591	0.0715	*0.0447*	0.0366	0.1804	0.3499	0.0667
	ERF	*0.4340*	*0.1560*	*0.0573*	0.0494	*0.0353*	*0.1229*	0.3080	0.0659
	EIIF	0.4822	0.1801	0.0621	0.0468	0.0354	0.1281	*0.2926*	0.0646
MEPV	CLONE	0.2671	0.0887	0.0523	0.0394	0.0454	0.0793	0.1600	0.0507
	PARENT	0.4046	0.1555	0.0671	*0.0410*	0.0438	0.1966	0.3355	*0.0649*
	ERF	*0.4044*	*0.1470*	*0.0581*	0.0538	0.0408	*0.1227*	0.3098	0.0661
	EIIF	0.4501	0.1862	0.0671	0.0465	*0.0363*	0.1348	*0.2671*	*0.0649*

Figures 6 to 8 clearly show, for the middle values of theta (-1, 0 to 1), the ERF, EIIF, and PARENT parameters had little discrepancy from the CLONE parameters and in some cases performed even slightly better than the CLONE parameters, as mentioned previously. However, for the lower or higher values of theta, the CLONE parameters had much smaller MSE values than the other three kinds of parameters. For the higher values of thetas, the PARENT parameters consistently performed the worst while, for the lower values of thetas, either PARENT or EIIF parameters performed worse across three item-selection methods. Note that there is one unexpected but interesting finding from Table 2: the PARENT parameters at theta = 0 performed better than the ERF and the EIIF parameters across three item-selection methods.

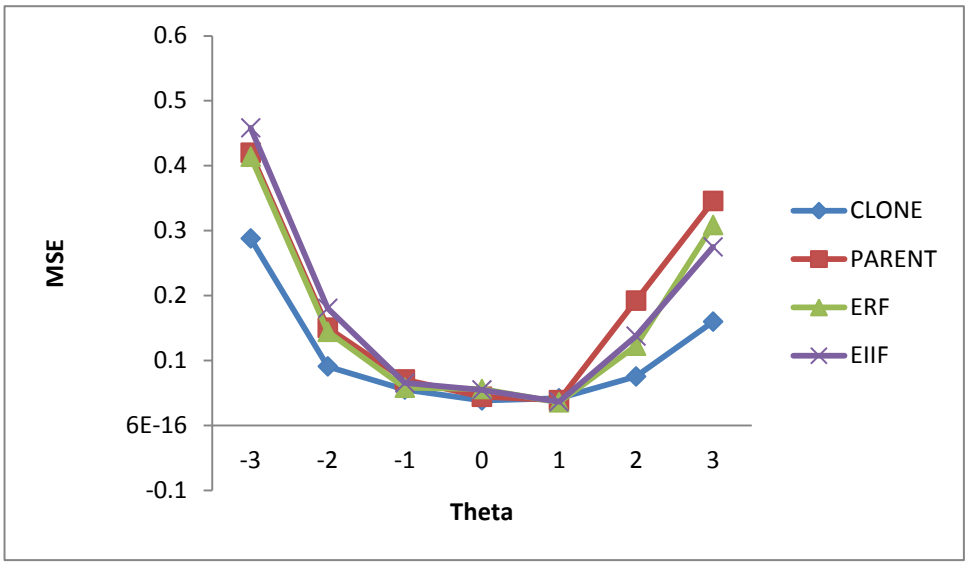

Figure 6. The MSE results of the MII method.

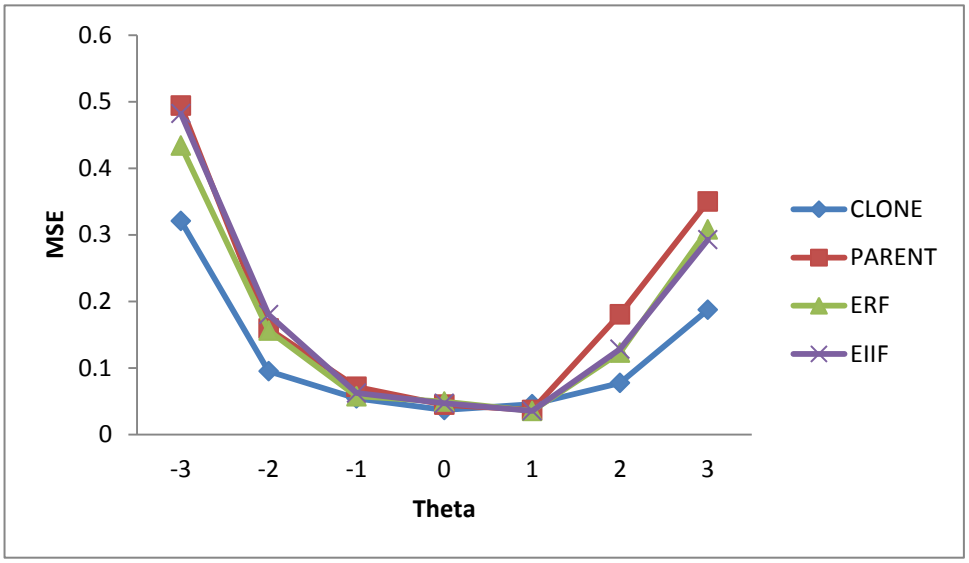

Figure 7. The MSE results of the MPWI method.

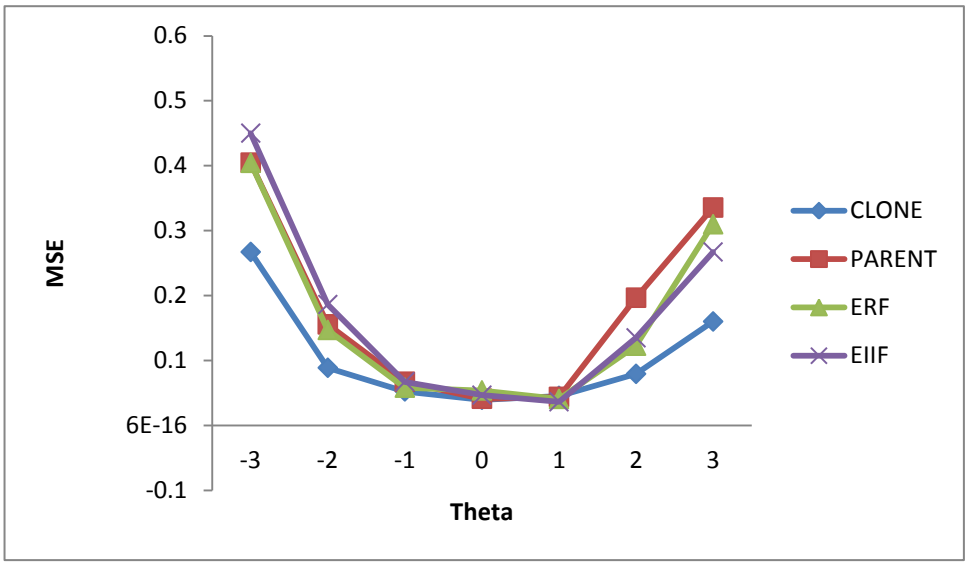

Figure 8. The MSE results of the MEPV method.

Conclusions and Further Research

There are two research questions to be answered through this study. First, is ERF or EIIF a good solution to model the variability within families in CAT? In general, the use of the ERF parameters was able to consistently reduce the possible precision loss to some extent, while using the EIIF parameters could sometimes perform worse than the PARENT parameters under some theta levels.

The second question is do different item-selection methods, including the maximum item information method, the MEPV method, and the MPWI method, make differences on measurement precision? The three item-selection methods produced similar results for each of four different types of parameters. Figure 9 shows the example of the MSE results of the ERF parameters for the three different methods.

The main findings of this study can be summarized as follows: first, when the items are sufficient for certain ability levels, the loss of precision is not observed or is small; and second,

the ERF parameters are promising in terms of reducing the possible precision loss due to the lack of isomorphic clones and the lack of adaptive item templates in the pool.

Further studies are needed. The following factors can be considered in further research: first, more pools should be examined with different levels of variability within templates; second, content balancing should be added to the CAT design; and third, exploring how many extra items are needed to compensate for the precision loss in variable-length CAT, if the precision loss exists.

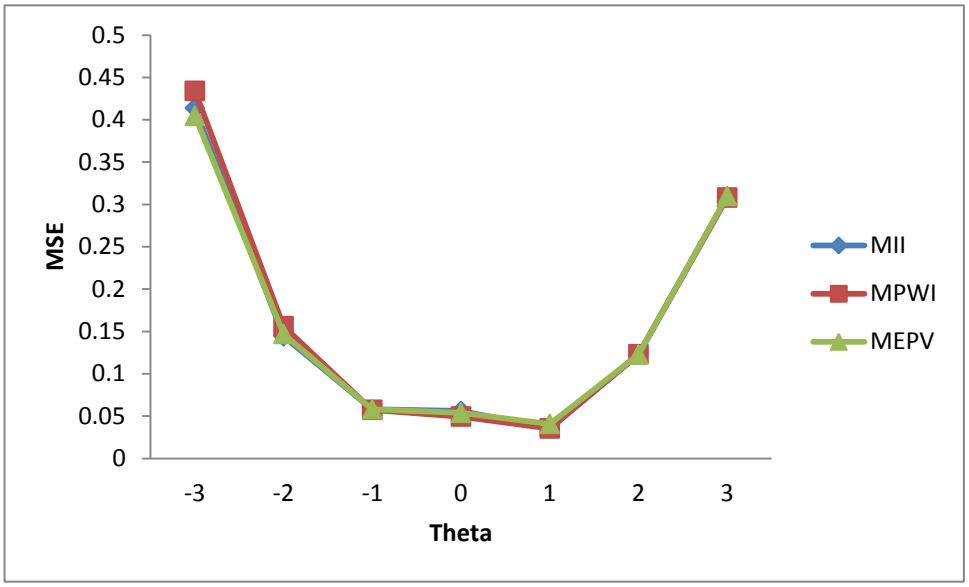

Figure 9. The MSE results of the ERF method for the three different item-selection methods.

References

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/viewFile/1663/1505

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation of item statistical characteristics. *Applied Measurement in Education*, *15*(1), 49–74.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 247–261.

Lewis, C. (1985, June). Estimating individual abilities with imperfectly known item response functions. Paper presented at the Annual Meeting of the Psychometric Society, Nashville, TN.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders. (Eds.), *Essays on Item Response Theory* (pp. 163–171). New York. NY: Springer

Meisner, R. M., Luecht, R., & Reckase, M. D. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms* (ACT Research Report Series No. 93-9). Iowa City, IA: The American College Testing Program.

Mislevy, R. J., Sheehan, K. M. & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Education Measurement, 30*(1). 55–78.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report RR-94-28-ONR). Princeton, N J: Educational Testing Service.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. Elements of Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Element of Adaptive Testing* (pp. 3–10). New York. NY: Springer

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