Methods for Item Set Selection in Adaptive Testing[1,2]

Ying Lu
University of Massachusetts Amherst

Saba Rizavi
Educational Testing Service

- April 13, 2003 -

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
[2] Center for Educational Assessment Research Report No. 487. Amherst, MA: University of Massachusetts, School of Education.

Introduction

Recent advances in technology have expedited the development of a more efficient testing format – computer adaptive testing (CAT), which tries to match item difficulty to the examinee's estimated ability at each step of the exam. When items are in individualized format, i.e., items are not bundled or grouped through their relationship to a common stimulus, CAT reaches its optimum efficiency. On the other hand, however, when items are in set format, e.g., when a number of items are based on a common reading passage in a verbal test, the level of adaptation that CAT can achieve is more limited. This is because whenever a set has been selected, CAT algorithm is restricted to selecting from the remaining items in the set, which may not be desirable, either statistically or cognitively.

Discussions started long ago over CAT algorithm designs aiming to alleviate the degree of inefficiency brought about by the special format of set items. The CAT algorithm for set-based items usually involves two steps repeatedly: deciding upon the set to be selected, and subsequently upon the items to be selected within the set. While there is not much adaptation possible in the latter step, an efficient algorithm for entering a new set becomes especially important. Intuitively, we would want to take into account set properties while making the decision over the set to be selected. Theunissen (1986, p.387) suggested that sets of items could be incorporated into a maximum information adaptive testing paradigm by using a set information function, which is the sum of the item information functions for the items within that set. Alternatively, the specific information item selection, in contrast with the maximum information selection, can be considered for the CAT of item sets. For a description of the specific information selection method, see Davey and Fan (2000). When applied to item sets in CAT, specific information selection method selects a new set that has expected set information values matched most closely to the target information value, where information value is obtained by numerically integrating the information function over the posterior ability estimate of the examinee. See Thompson & Davey (1999, 2000) for the application of specific information item selection to a passage-based test. The problem with using expected set information function or set information values is that very often the number of items to be administered from each set is not known in advance, and even when it is known, it usually differs from the number of items available in the set. Besides, with other complicating factors, e.g., mixed item format, and large number of test construction rules both at the set level and the item level, the efficiency of incorporating expected set information function/value into the algorithm is not known.

Stocking & Swanson (1993) proposed the use of a weighted deviations model (WDM) and an algorithm for item selection. Particularly in dealing with the item set format, they suggested assigning a conceptual partition of the item pool (a block) to each set, and entering a block by the selection of an item in that block that contributes the most to the satisfaction of all other constraints (p. 282). Once within a block, items continue to be selected adaptively for administration based on their contribution to the satisfaction of all constraints and overlap, until the number of items to be administered from that block is reached. This is the item set selection method most testing programs are implementing now. The weighted deviations model has been found to be quite satisfactory in its capability of handling large number of constraints on intrinsic item features. The efficiency of this methodology, in dealing with set-based items, however, is threatened by the negligence of set attributes at the selection of a set, as set selection is determined solely by the characteristics of one single item. It is possible that, after an examinee has been administered the first item of a set, none of the remaining items in the set match the examinee's interim ability estimate well. As a result, items are administered without being able to obtain useful information about the examinee.

Theoretically, if item sets with different item difficulty distributions could be administered at different stages of CAT administration, it would allow for more efficiency under this special item format condition. At earlier stages when the ability estimates for the examinees have considerably large standard errors, a set of items with heterogeneous difficulties is preferred as it leaves space for estimation errors and provides more flexibility. At later stages when the ability estimates become increasingly accurate, a set of items with homogeneous difficulties is preferred as it improves precision within a fine range of ability estimates. The purpose of this paper is to examine the effect over ability estimation accuracy of making variations to the WDM to take into consideration the item difficulty distributions within sets for tests comprised primarily of item sets.

Method

Conventional test construction usually needs to take into account test specifications. And accordingly, various rules are built into the CAT algorithm to constrain the automatic item selection process. Stocking and Swanson (1993, p. 278) put test specifications into four categories: (1) constraints on intrinsic item properties; (2) constraints on item features in relation to all other candidate items (overlap); (3) constraints on item features in relation to a subset of all other candidate items (item set); (4) constraints on the statistical properties of items as derived from pretesting.

Below is a brief review of how constraints are built into WDM. For a detailed introduction, readers are referred to Swanson, L., & Stocking, M. L. (1993), Stocking and Swanson (1993).

The basic form of WDM is intended to find the item that minimizes the function of weighted sum of deviations (WSD):

$$\sum_{j=1}^{J} w_j d_{L_j} + \sum_{j=1}^{J} w_j d_{U_j} + w_\theta d_\theta , \qquad (1)$$

where $j = 1, \ldots, J$ indexes the item properties associated with the content constraints, w_j denotes the weight assigned to constraint *j*, w_θ denotes the weight assigned to the information constraint, d_{L_j} and d_{U_j} denote the non-negative deviations from the lower and upper bounds respectively for content constraint *j* if the item is selected, and d_θ denotes the non-negative deviation from the lower bound on test information if the item is selected.

This weighted deviation model where the weighted sum of deviations is defined by equation (1) is used as the baseline method for item and set selection in this paper, which is implemented in simulation study one. In this method, a set is selected due to a single item within the set that has desirable content and information properties, as demonstrated by smallest weighted sum of deviations defined in equation (1). After a set is selected, items continue to be selected adaptively within the set. Here, we make the assumption that item review by examinees is not allowed, thus making adaptive testing within a set possible.

In second and third simulation studies, we build into the WDM the extra constraint concerning item difficulty distributions within sets, to enhance the model's capability to deal with statistical properties in the context of set format. The extra constraint is intended to minimize the difference between standard deviation of item difficulties within a set and the standard error of the interim ability estimate. The standard error of the interim ability estimate for examinee *i* at the *r*th CAT stage (Hambleton, Swaminathan & Rogers, 1991) is known to converge to:

$$SE\left(\hat{\theta}_{ir}\right)=\frac{1}{\sqrt{I\left(\hat{\theta}_{ir}\right)}},$$

where $\hat{\theta}_{ir}$ is the interim θ estimate for examinee *i* after the preceding *r*-1 items, and $I\left(\hat{\theta}_{ir}\right)$ is the sum of the item information functions at $\hat{\theta}_i$ for the preceding *r*-1 items.

Let $SD_s\left(b\right)$ denote the standard deviation of item difficulties within set *s*. The constraint of minimizing $\left|SE\left(\hat{\theta}_{ir}\right)-SD_s\left(b\right)\right|$ is intended to take into account the uncertainty of the ability estimate when selecting a set. For this constraint there is no lower and upper bounds. The closer the value of $\left|SE\left(\hat{\theta}_{ir}\right)-SD_s\left(b\right)\right|$ approaches 0, the better the item set is matched to the interim adaptive testing stage. Now, let d_{SD} denote the value of $\left|SE\left(\hat{\theta}_{ir}\right)-SD_s\left(b\right)\right|$, and let w_{SD} denote the weight assigned to the new constraint. The modified weighted sum of deviations is defined as

$$\sum_{j=1}^{J}w_jd_{L_j}+\sum_{j=1}^{J}w_jd_{U_j}+w_\theta d_\theta+w_{SD}d_{SD} \qquad (2)$$

The minimization of this weighted sum of deviations can be used as a criterion for selecting items to be administered. When the selection of an item determines the new block of item set to be entered, this minimization function takes into account item set difficulty distribution to allow for the maximum efficiency. When items are selected adaptively within a set, $w_{SD}\, d_{SD}$ serves as a dummy function that does not play any role in the item selection process, as it takes the same value for items within the same set.

This variation of WDM is used in the second simulation study. The procedure followed for set selection in this study is similar to that in the first study, except that desirable item properties now include the compatibility of the item difficulty distribution within the set to which the item belongs with the SE of the interim ability estimate. In other words, $w_{SD}d_{SD}$ is integrated as a term in the weighted sum of deviation.

The third simulation study utilizes more set properties including the expected set content attributes and expected set information while selecting a set. This is an adaptation of Theunissen's suggestion of using a set information function, and is only viable with the assumption that the number of items to be administered from each set is known in advance, the expected set information at $\hat{\theta}_i$ is computed as

$$E\left(I_s\left(\hat{\theta}_i\right)\right)=\frac{m_s}{n_s}\sum_{k=1}^{n_s} I_k\left(\hat{\theta}_i\right),$$

where n_s is the number of available items in set *s*, m_s is the number of items to be administered from set *s*, $I_k\left(\hat{\theta}_i\right)$ is the item information for the *k*th item in set *s*. Similarly, the expected set content attribute is computed by multiplying the sum of item content attributes within the set by m_s / n_s.

In this study, when a new set of items needs to be administered, the function of weighted sum of deviations is redefined over sets instead of items. The form of the function stays the same, but the parameters in it take different denotations:

$$\sum_{j=1}^{J} w_j d_{L_j} + \sum_{j=1}^{J} w_j d_{U_j} + w_\theta d_\theta + w_{SD} d_{SD}$$

where d_{L_j} and d_{U_j} denote the expected non-negative deviations from the lower and upper bounds for content constraint *j* when the particular SET is selected, and d_θ denotes the expected deviation from the lower bound on test information when the SET is selected. A set is selected when it has the smallest set-based WSD. After the set to be administered is determined, items are selected adaptively within the set, based on the item-level WDM.

Simulation Design

The simulations utilize an item pool of 438 dichotomous items (64 sets with at least 6 items in each set) from several paper and pencil forms of a large-scale test comprised of passage related sets. Item parameters were obtained from PARSCALE calibration using three-parameter logistic model. These item parameters are treated as true parameters, which is a realistic assumption since they were estimated from a considerably large sample of examinees, and in most CAT situations this is the case. The decision of using real test data against simulated item parameters for the item bank was made so that the study would be operationally based, and with realistic significance. Table 1 summarizes the means and standard deviations of the item parameter estimates for the pool. Figure 1 displays the histogram of *b*-parameters in the Pool. The table and the figure show that the distribution of *b*-parameters of items in the pool is slightly negatively skewed, with the mode around 0.0, but mean at –0.31. The guessing parameters are relatively high with mean at 0.28.

Insert Table 1 about here

Insert Figure 1 about here

Figure 2 shows the histogram of the within-set item difficulty standard deviations for all the 64 sets in the pool. It is noticed that the SDs have a pretty condensed distribution, where most values gather between 0.35 to 1.40.

Insert Figure 2 about here

A fixed length of 35 questions (seven, five-item sets) is determined for the adaptive test. To reflect typical assessment where the balance between psychometric properties and content specifications need to be met, the simulation studies take into account content specifications. Table 2 lists the six content specifications used in the simulations. Content specifications are set in terms of minimum and maximum number of items per content category, and proportion of test items within each content category for the fixed test length of 35 questions, which, in Table 2, is denoted as target proportion. The chosen content specifications are so devised to approximate those used in practice.

Insert Table 2 about here

Content attributes for items in the pool are generated according to the target proportion, under the assumption that attributes are independent across content categories. For a particular content category and for a specific item, a random number is drawn from uniform distribution with intervals 0 to 1. If this number is larger than the target proportion for this content category, the content attribute for the item is set to 0; if it is smaller than the target proportion, the content attribute is set to 1. Content attributes are so simulated based on the belief that tests are condensed forms of the pool, and mean content attributes for items in the pool should more or less reflect the target proportion. After the content attributes are simulated for all the items, proportions of test items in the pool having the specific content attributes are computed and summarized also in Table 2. An equal weight of 1.0 is assigned to all the content constraints showing that content categories are equally important. In practice, however, the weights might be different for different content constraints depending on the content and psychometric experts' judgment.

Data for 500 simulees are generated at 15 ability levels, resulting in a total of 7,500 simulees. The ability levels are unequally spaced to approximate the *N* (0, 1) target population distribution (Robin, 2001): -1.93, -1.28, -0.96, -0.72, -0.52, -0.33, -0.16, 0.0, 0.16, 0.33, 0.52, 0.72, 0.96, 1.28 and 1.93. The 500 replications at each ability level are judged to be necessary for obtaining stable estimates of conditional results.

For examinee ability estimation, expected a posterior (EAP) is used because of its capacity to produce estimates for candidates who score the highest on all items, or the lowest on all items. In dichotomous CATs, the EAP estimates have been shown to produce smaller mean square error terms over the population than MLE or MAP (Chen, et al, 1997). To avoid estimation bias, a weak prior is used, with mean set at 0, and SD set at 2.0.

To insure fair comparisons among the three set selection methods, which is to say, no method should be found superior than the other method in terms of measurement precision because less importance is attached to the satisfaction of content constraints, the proportion of total weights assigned to the content constraints is fixed across the three methods. Initial simulations have been run to examine the effect of different partitions of WDM weights between content constraints and item information over measurement precision and satisfaction of content specifications. It is found that with a weight of 25.0 assigned to item information, and a total weight of 6.0 assigned to the content constraints (1.0 for each content category), there is a reasonable balance between measurement precision and content constraint violations (see result section). Therefore, this assignment of weights is used in the baseline simulation study. In the second and third simulation studies, the weight of non-content constraints, which in study one is all assigned to item information, would be shared between information function and d_{SD}, which means

$$\sum_{j=1}^{6} w_j = 6,$$

$$w_\theta + w_{SD} = 25.$$

Within the 2nd and 3rd set selection methods, simulations are conducted using several partitions of weight of non-content constraints into w_θ and w_{SD}. Combinations include 21/4 (w_θ / w_{SD}), 17/8, 13/12, 9/16. Details of WDM weight assignment in this study are summarized in Table 3.

Insert Table 3 about here

Simulations were run using SETSIM (Lu & Robin, 2003), a program developed for CAT simulation of set-based items. Simulation results are evaluated according to measurement and content criteria. Evaluation of content is based on the extent to which content specifications are satisfied. Measurement results are evaluated in terms of conditional bias (*CB*), conditional standard error of measurement (*CSEM*), and conditional root mean squared error (*CRMSE*). Let *r* index the replication, where *r* = 1,

…, 500. Let $\bar{\hat{\theta}}$ denote the mean estimated ability across replications for the true ability level θ. *CB*, *CSEM*, and *CRMSE* are computed at 15 levels of θ

$$CB\left(\theta\right)=\theta-\bar{\hat{\theta}}\ ,$$

$$CSEM\left(\theta\right)=\sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\theta}_r-\bar{\hat{\theta}}\right)^2}\ ,$$

$$CRMSE\left(\theta\right)=\sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\theta}_r-\theta\right)^2}\ .$$

These three evaluation criteria are related through $CRMSE^2=CB^2+CSEM^2$.

Results

The average percentage of content constraint violations is reported for each content category for the baseline simulation (See Figure 2) to examine whether content aspects are given adequate consideration. The average percentages of violations range from 0.104 to 0.342, with the mean being 0.237, which is regarded as quite reasonable. The percentages are relatively small for constraints 1, 2 and 3, and relatively large for constraints 4, 5 and 6. This difference in content violation percentages is mostly due to the difference in the interval lengths for satisfactory content attributes, i.e., the difference between lower and upper bounds. The proportion of violations tends to be small, when there is a larger interval and satisfaction of this particular content specification is defined on a more flexible basis.

Insert Figure 3 about here

Other simulation studies have produced similar content results, as the proportion of weights assigned to content constraints is the same across all simulations. Therefore, we will not elaborate further on the content results. Assuming content specifications are reasonably satisfied with all simulation studies, we will compare measurement precisions across simulations.

Measurement results are summarized in terms of *CB*, *CSEM*, and *CRMSE*, and are compared across the three simulation designs for different WDM weights allotment.

Figures 3a to 3c display the conditional measurement results for methods 1, 2 and 3, with a weight of 21 assigned to information and a weight of 4 assigned to compatibility of within set difficulty distribution and SE of interim ability estimate in method 2 and 3. As can be seen in Figures 3a, across the three simulations, there is small positive bias before 0.96 on the true ability scale, and negative bias after that. This is mostly due to the $N(0, 2.0)$ prior distribution used in the EAP estimation. Method 3 has slightly larger biases than method 1 and 2. With regard to *CSME* and *CRMSE*, method 3 gives similar performance with method 1 and 2 at the middle range of ability levels, but less good performance at the lower and upper end of the ability scale. Method 1 and 2 are giving similar results.

Insert Figures 4a to 4c about here

Figures 4a to 4c display the conditional measurement results for methods 1, 2 and 3, with a weight of 17 assigned to information and a weight of 8 assigned to compatibility of within set difficulty distribution and SE of interim ability estimate in method 2 and 3. Note that compared with the previous simulation conditions, w_θ has decreased by 4.0, while w_{SD} has increased by 4.0. Still, method 3 has slightly larger biases than method 1 and 2. With regard to *CSME* and *CRMSE*, the three methods are giving more similar results compared to those in the previous simulation conditions. Noticeably, method 2 is giving slightly better results above ability level of 0.16, and method 3 is performing least well at the lower and upper end of the ability scale.

Insert Figures 5a to 5c about here

Figures 5a to 5c display the conditional measurement results for methods 1, 2 and 3, with a weight of 13 assigned to information and a weight of 12 assigned to compatibility of within set difficulty distribution and SE of interim ability estimate in method 2 and 3. In this simulation study, information and within-set *b* distribution are almost taking the same weight. As can be seen, method 1 performs slightly better than the other 2 methods. Method 2 has lost the edge it has shown in the 2nd combination of weights, and method 3 performs less well at the two ends of the ability scale. But still, the difference among the three methods is small.

Insert Figures 6a to 6c about here

Figures 6a to 6c display the conditional measurement results for methods 1, 2 and 3, with a weight of 9 assigned to information and a weight of 16 assigned to compatibility of within set difficulty distribution and SE of interim ability estimate in method 2 and 3. As expected, since the weights for information are considerably reduced in method 2 and 3, which cannot be compensated by the increase in the amount of weight to d_{SD}, method 1 is consistently giving better *CB*, *CSME* and *CRMSE* results along all ability levels. It is also interesting to note that in this particular combination of weights, method 3 is performing better than method 2.

Insert Figures 7a to 7c about here

Discussion and Conclusion

Since the special feature of set format allows for only partial adaptation, the constructed test comprised primarily of set-based items may not have the desired measurement properties. Whenever a set has been selected, the CAT algorithm is restricted to selecting the remaining items in the set, which may not be desirable, either statistically or cognitively. It is believed that the problem of having to use statistically dissatisfactory items is exacerbated when a set is selected based on a single item within the set that is considered the most desirable. The problem is partly due to the neglecting of set properties when the decision on set selection is being made.

One of the set properties that simulations in this paper have included is the within-set item difficulty standard deviation. However, how much adding the constraint that selects sets with SD of difficulties matched to the SE of interim ability estimate could alleviate the inefficiency problem depends on many factors.

One factor is the weight assigned to this extra constraint. In this study, we have found that neither too much nor too little weight would help improve measurement precision. Too much weight puts the information function, which is the most important determinant of measurement error, at a less important position, and doing so greatly threatens the efficiency of CAT. Even though the extra constraint is expected to play a significant role in measurement precision, its role is far less important than the information function, and therefore, should be given correspondingly less weight. At the same time, too little weight is not effective to the point that the extra constraint would be able to have a significant impact on set selection. As is demonstrated in the study in this paper, of the four weight combinations, the second method gives the best results in the weight combination of 17 assigned to information and 8 assigned to the extra constraint, although the edge is small. Consequently, in testing practice, when it is decided that the extra constraint is to be used, simulation studies need to be conducted to identify the most efficient allotment of weights among content specifications, information, and d_{SD} to achieve the balance of content specification satisfaction and measurement precision.

The second factor affecting the use of the extra constraint is the distribution of within-set item difficulty *SD*s. When the variance is small, which is to say, there is not much variability among the $SD_s\left(b\right)$'s, there is little sense adding this extra constraint as it will not differentiate sets. A follow-up simulation study that would be of interest is to use simulated item parameters with $SD_s\left(b\right)$ being a varying factor. It would be of practical significance to identify the borderline of the variance of $SD_s\left(b\right)$'s, above which adding the extra constraint would make significant improvement in ability estimation, and below which it would not.

Another interesting finding of this paper is about the performance of method 3. Method 3 is taking into account in CAT administration more set properties than the other two methods. However, our results have found that utilizing expected set information function while deciding on a new set to enter does not significantly improve measurement precision. On the contrary, it gives slightly larger *BIAS* and *CMSE* at the lower and upper ends of the ability scale. Giving a second thought on it, the value of the expected set information being high at $\hat{\theta}$, on most occasions, means that the items within the set have the most overlap of item information at $\hat{\theta}$, and that their *b* values average pretty much around $\hat{\theta}$. However, an average *b* value around $\hat{\theta}$ does not necessarily mean that items within the set can differentiate well around $\hat{\theta}$, as neither items with *b* values considerably above $\hat{\theta}$ nor items with *b* values considerably below $\hat{\theta}$ could well serve to give refined estimate around $\hat{\theta}$, yet they could comprise a set producing relatively high information at $\hat{\theta}$.

This study is critical for CAT administration of tests comprised of primarily item sets, where full adaptation of item selection is not possible. Besides collecting more empirical evidence about the effectiveness of utilizing the within-set item difficulty SDs, future research could also compare the methods used in this study with the specific information item selection method, which is also a promising research direction for CAT of set based items.

Reference

Chen, S., Hou, L., Fitzpatrick, S. I. & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing. *Educational and Psychological Measurement, 57(3),* 422-439.

Davey, T., & Fan, M. (2000, April). *Specific information item selection for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Lu, Y. & Robin, F. (2003) *SETSIM* [computer program]. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluation Research.

Robin, F. (2001). *Development and evaluation of test assembly procedures for computerized adaptive testing.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151-166.

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381-389.

Thompson, T. D. & Davey T. (1999, April). *CAT Procedures for Passage-based sets.* Paper presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Thompson, T. D. & Davey T. (2000, April). *Applying Specific Information Item Selection to a Passage-Based Test.* Paper presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Table 1. Item Parameter Estimates for the Item Pool for Simulations

# Items	#Sets	*a*-Parameters		*b*-Parameters		*c*-Parameters	
		Mean	SD	Mean	SD	Mean	SD
438	64	0.7	0.26	-0.31	1.19	0.28	0.08

Table 2. Content Constraints and Pool Content Attributes

Content	WDM Specifications			Target	Pool
Area	LB*	UB*	Weight	Proportion	Proportion
1	8	12	1.0	0.29	0.28
2	4	8	1.0	0.17	0.16
3	7	10	1.0	0.24	0.24
4	6	9	1.0	0.21	0.19
5	5	7	1.0	0.17	0.21
6	10	12	1.0	0.31	0.29

* Lower and upper bounds.

Table 3. WDM Weights in Simulation Studies

		WDM Weights		
		Content	Information	d_{SD}
Method 1		6	25	N/A
Method 2	comb. 1	6	21	4
	comb. 2	6	17	8
	comb. 3	6	13	12
	comb. 4	6	9	16
Method 3	comb. 1	6	21	4
	comb. 2	6	17	8
	comb. 3	6	13	12
	comb. 4	6	9	16

Figure 1: Histogram of *b*-parameters in the Pool

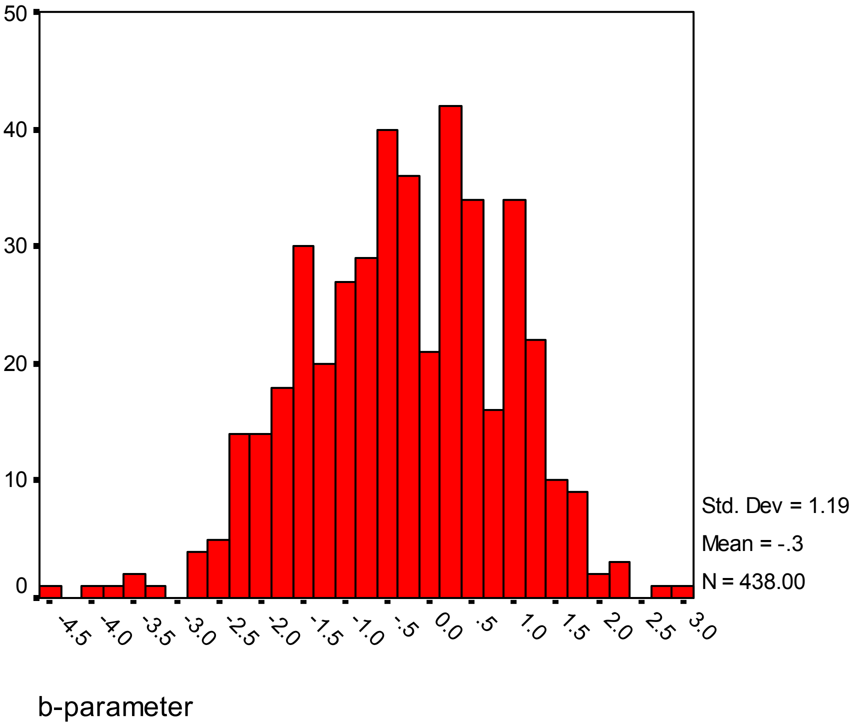


Figure 2. Histogram of the Within-set Item Difficulty SDs for Sets in the Pool

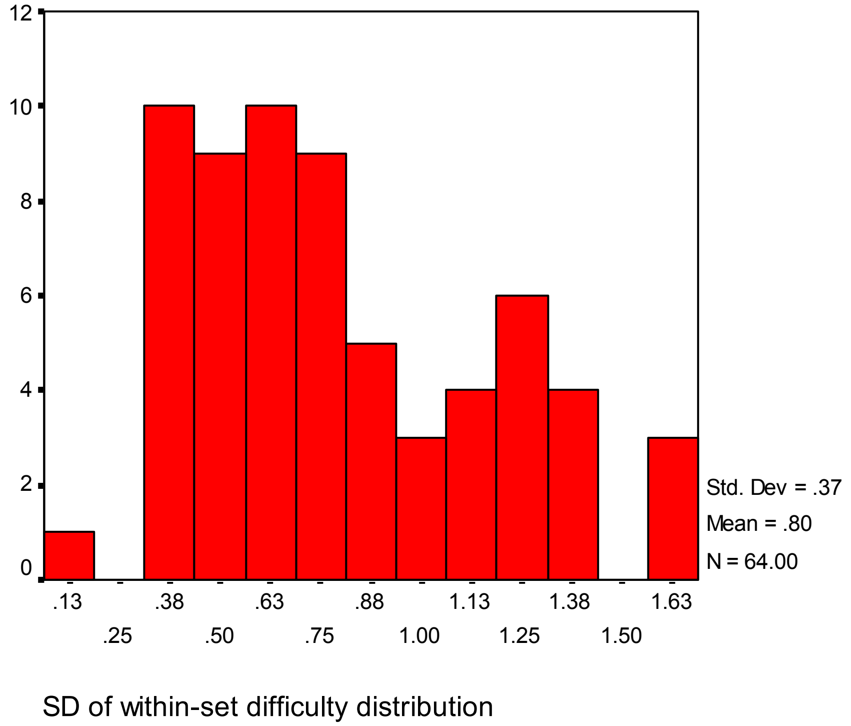


Figure 3. Proportion of Content Constraint Violations for the Six Content Categories

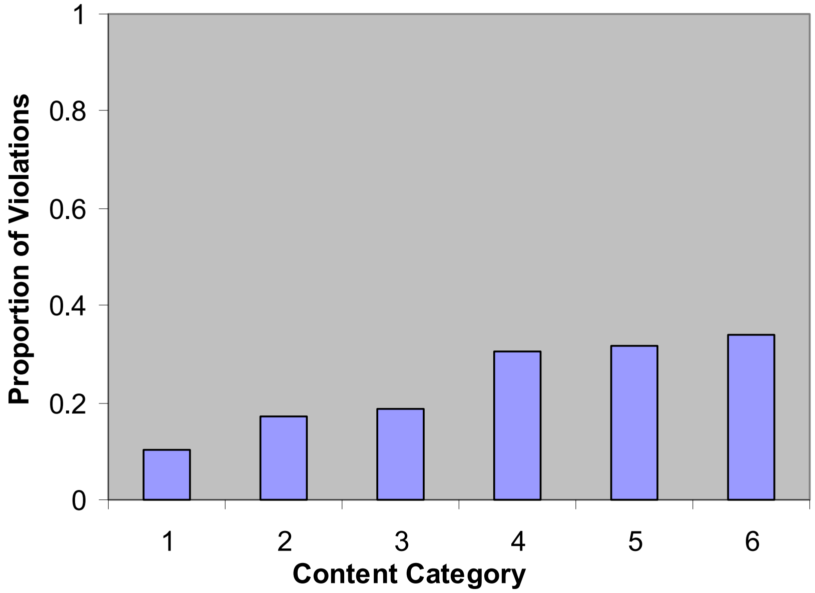


Figure 4a. Conditional Bias across the Three Methods with $w_{\theta} = 21$ and $w_{SD} = 4$

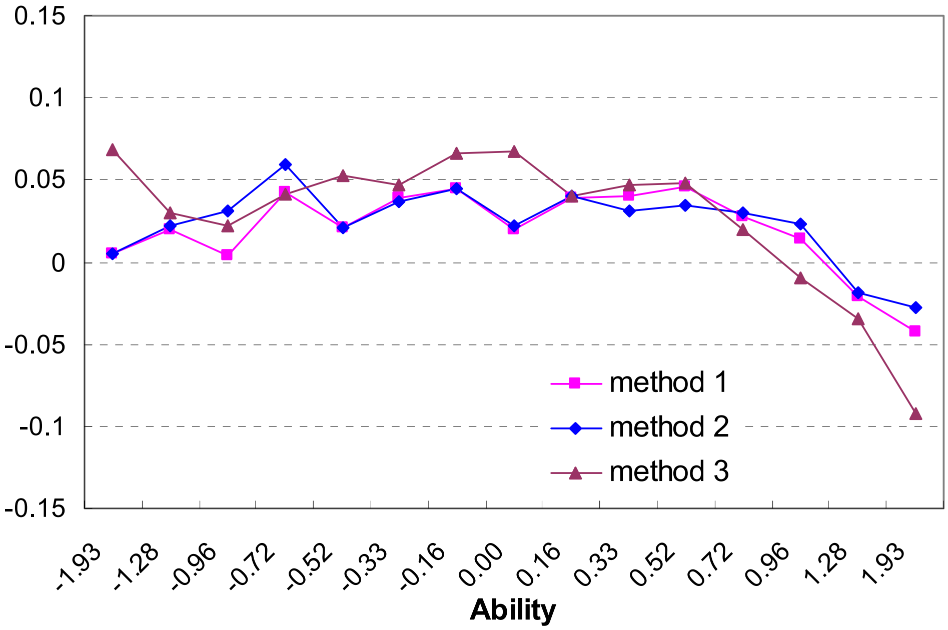


Figure 4b. Conditional Standard Measurement Error across the Three Methods with $w_{\theta} = 21$ and $w_{SD} = 4$

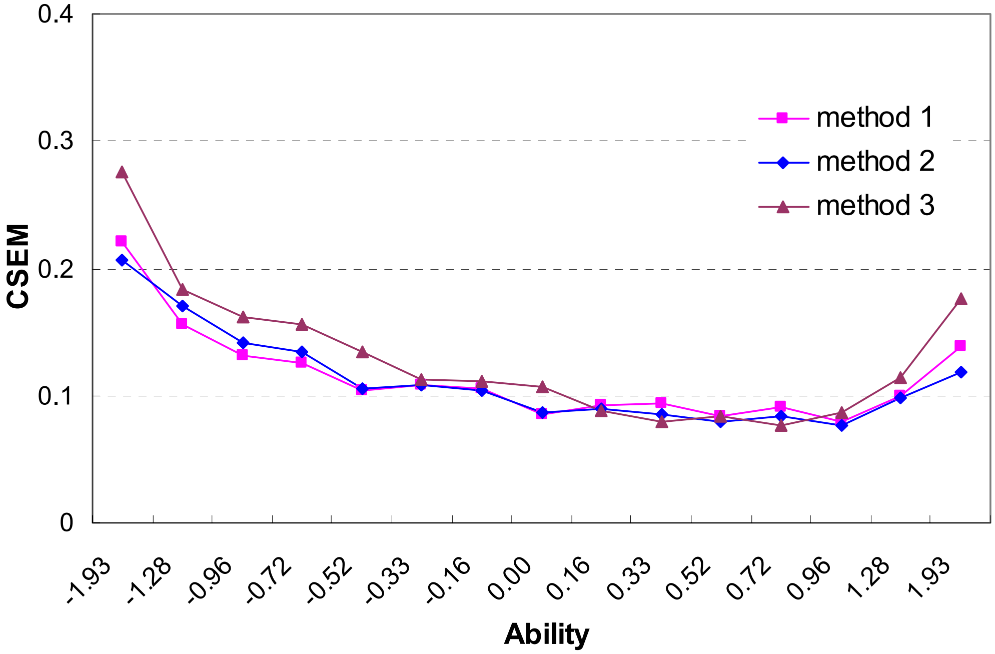


Figure 4c. Conditional Root Mean Squared Error across the Three Methods with $w_\theta = 21$ and $w_{SD} = 4$

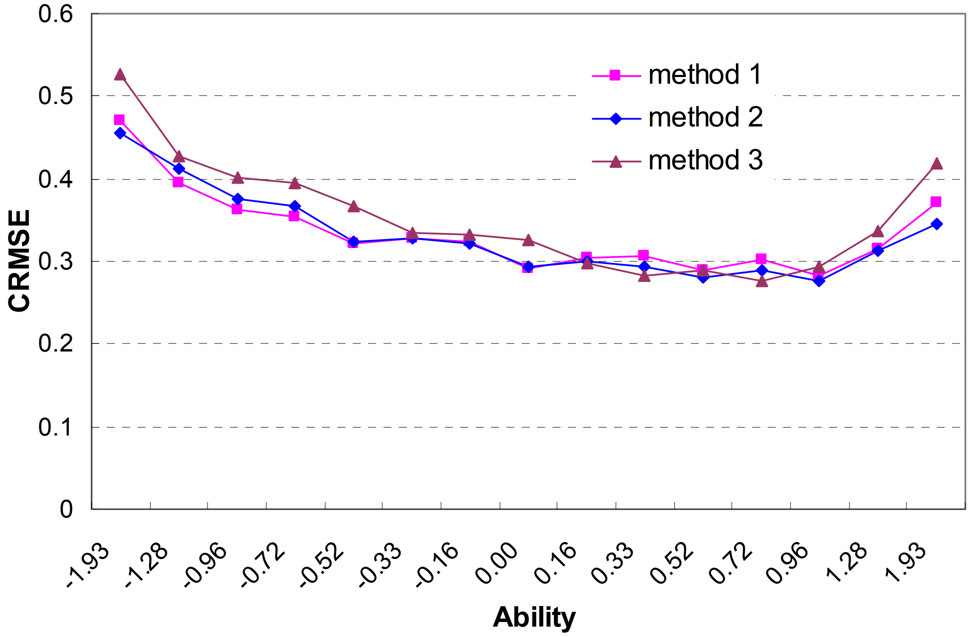


Figure 5a. Conditional Bias across the Three Methods with $w_\theta = 17$ and $w_{SD} = 8$

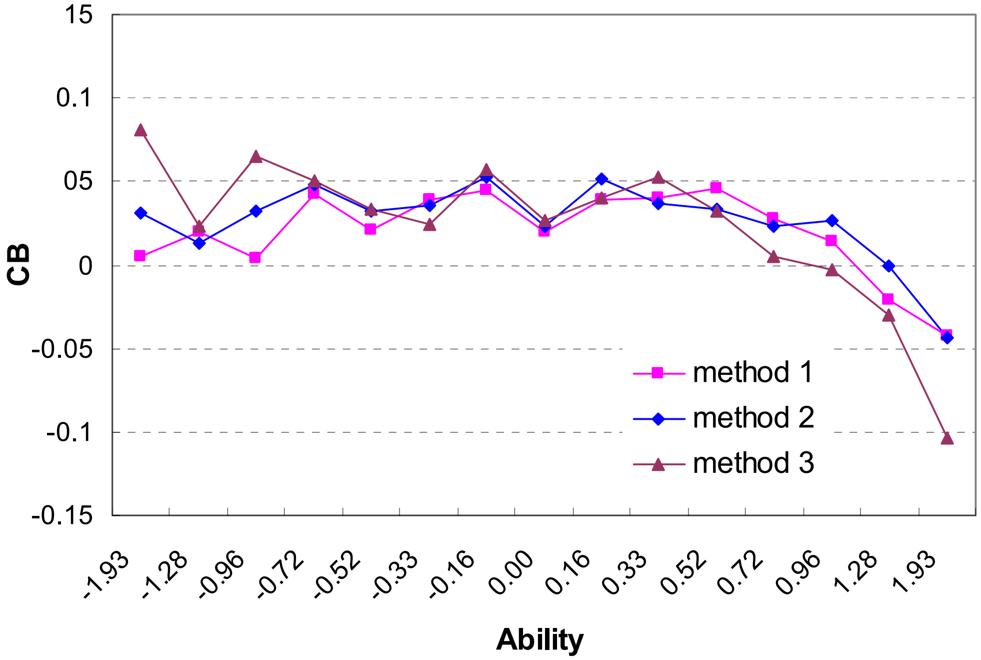


Figure 5b. Conditional Standard Measurement Error across the Three Methods with $w_\theta = 17$ and $w_{SD} = 8$

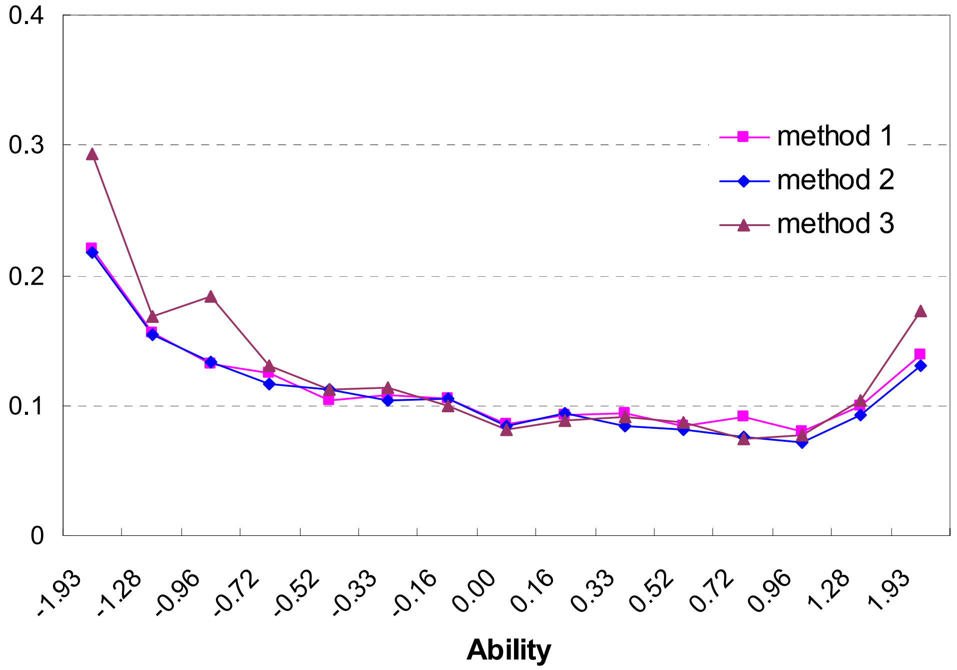


Figure 5c. Conditional Root Mean Squared Error across the Three Methods with $w_\theta = 17$ and $w_{SD} = 8$

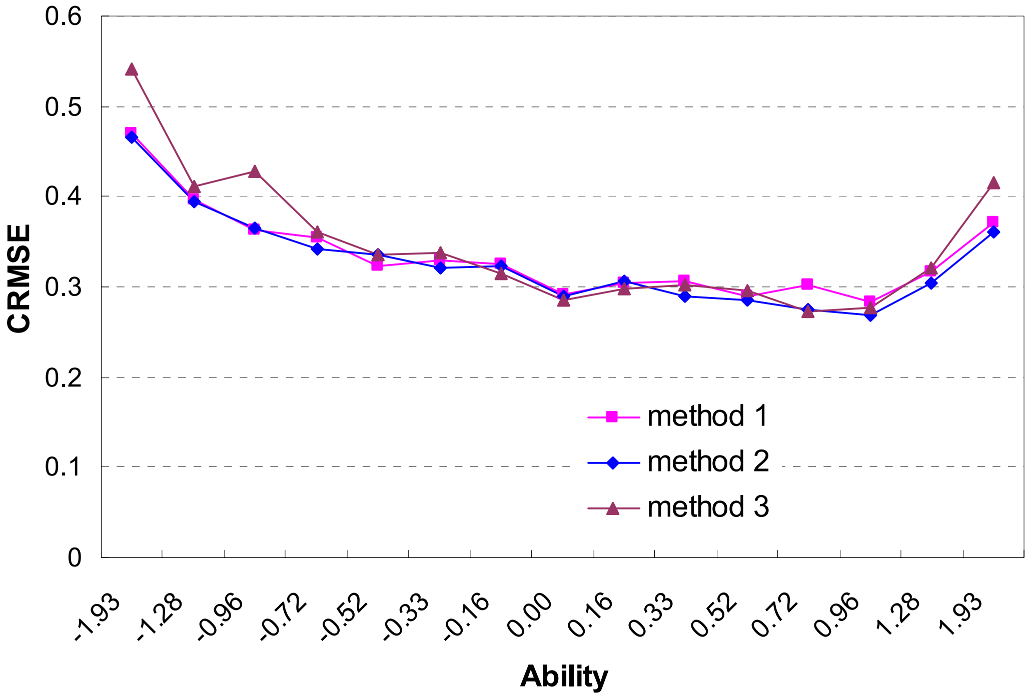


Figure 6a: Conditional Bias across the three methods with $w_{\theta} = 13$ and $w_{SD} = 12$

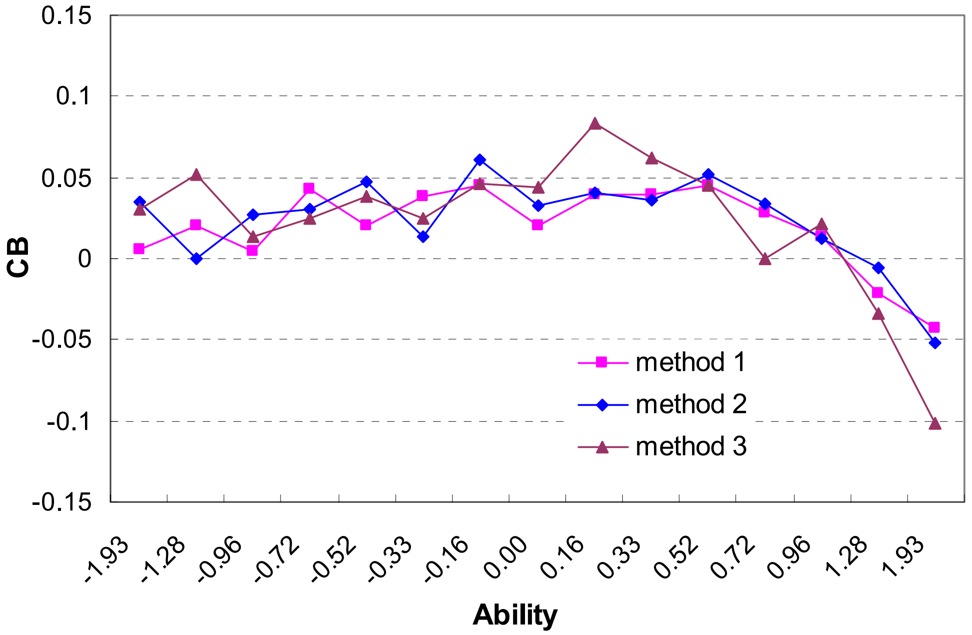


Figure 6b: Conditional Standard Measurement Error across the Three Methods with $w_{\theta} = 13$ and $w_{SD} = 12$

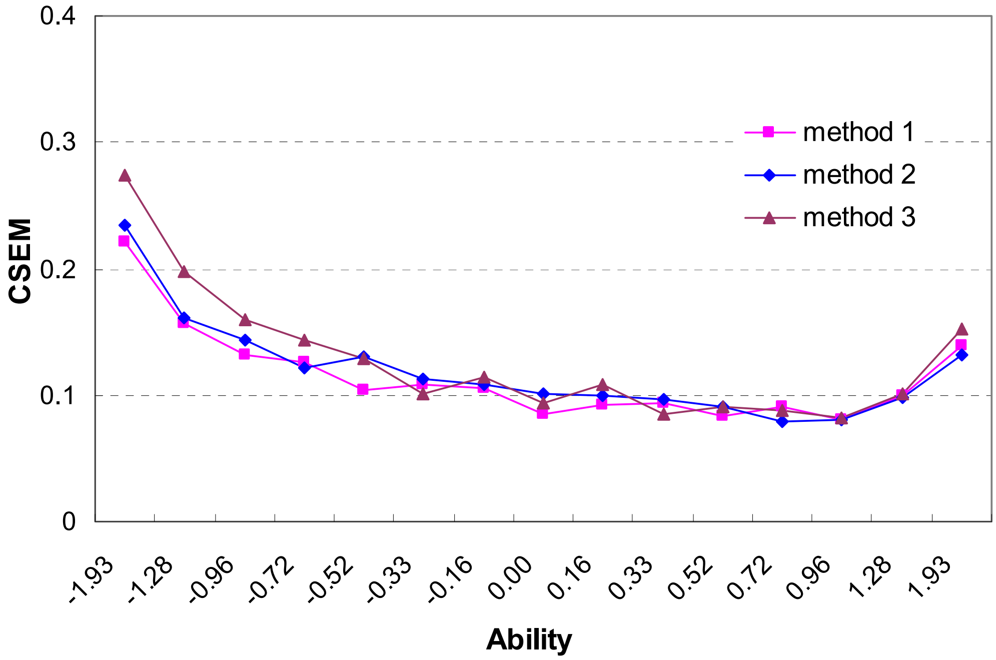


Figure 6c: Conditional Root Mean Squared Error across the Three Methods with $w_\theta = 13$ and $w_{SD} = 12$

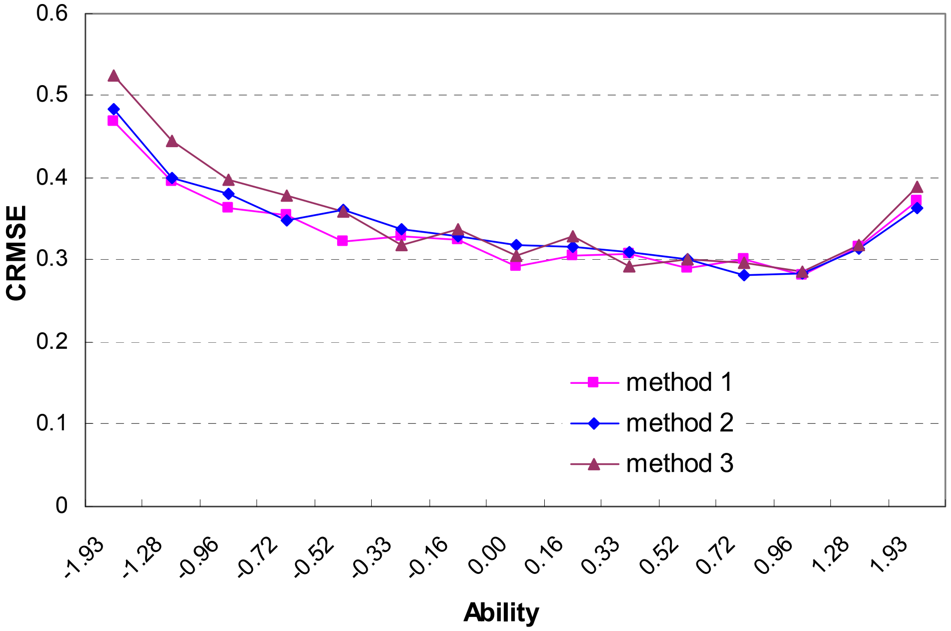


Figure 7a: Conditional Bias across the three methods with $w_\theta = 9$ and $w_{SD} = 15$

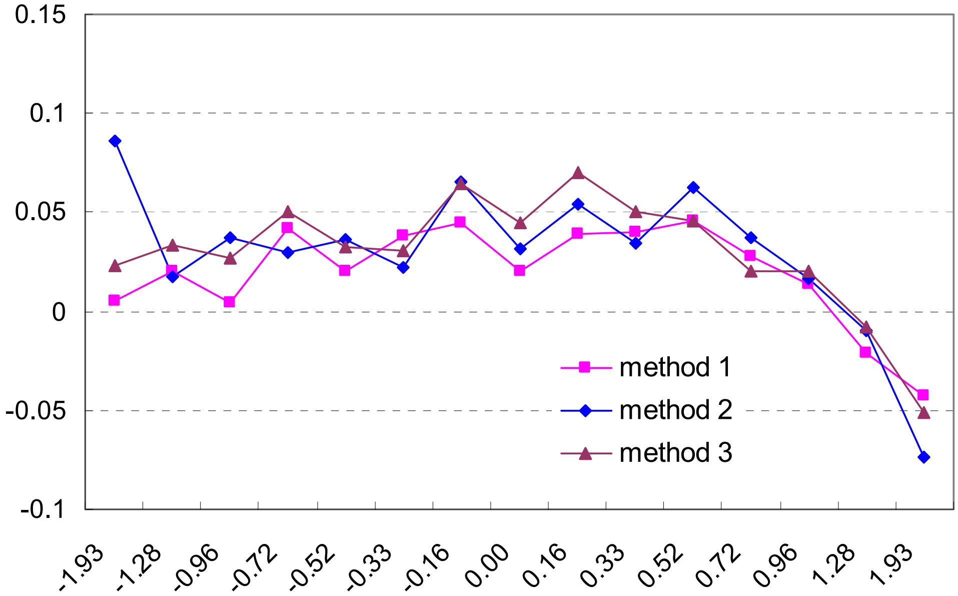


Figure 7b: Conditional Standard Measurement Error across the Three Methods with $w_{\theta} = 9$ and $w_{SD} = 16$

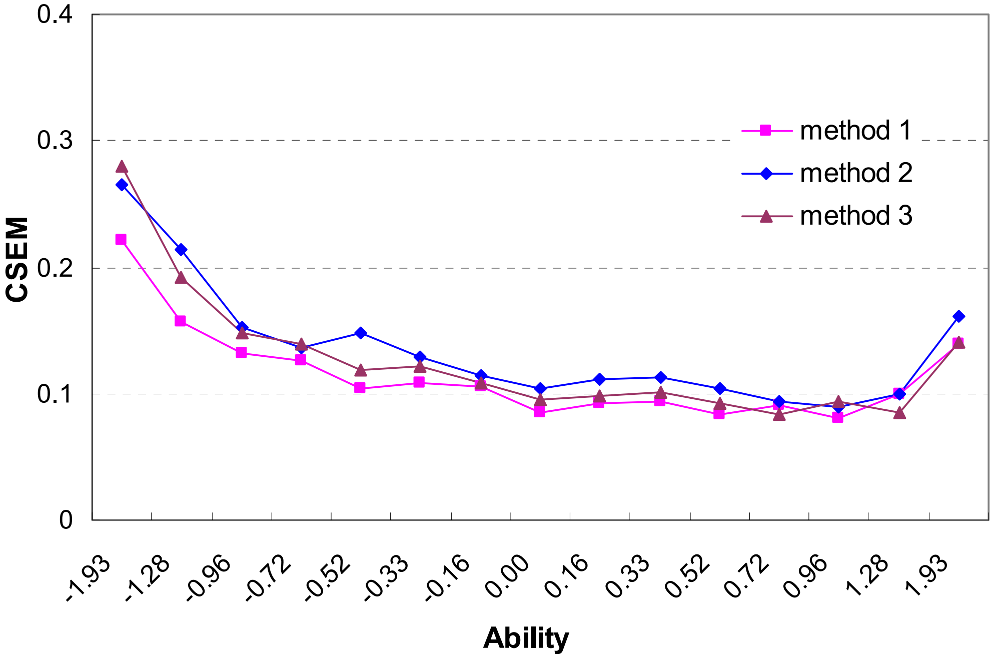
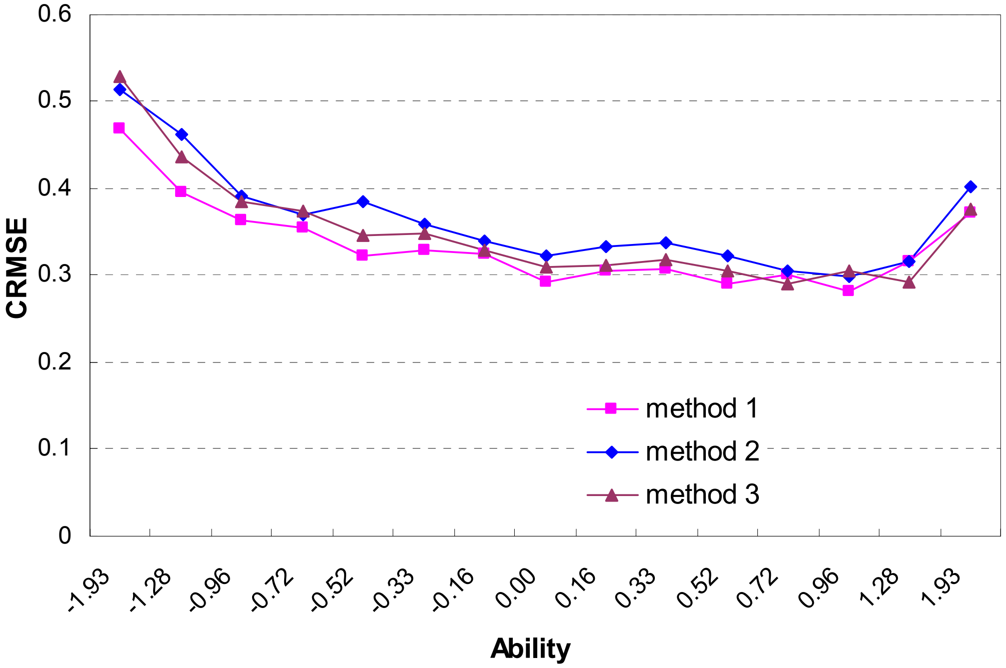


Figure 7c: Conditional Root Mean Squared Error across the Three Methods with $w_{\theta} = 9$ and $w_{SD} = 16$


CRMSE
method 1
method 2
method 3
Ability