Effects of Nonequivalence of Item Pools on Ability Estimates in CAT

Jae-Chun Ban, Tianyou Wang, Qing Yi, & Deborah J. Harris

ACT, Inc.

Paper presented at the annual meeting of the National Council on Educational Measurement, April 25-27, 2000, New Orleans, LA

ABSTRACT

This study focused on the effects of nonparallel item pools on examinees' ability estimates and on the number of administered items in each content category. Two sets of five nonparallel multiple item pools were constructed to be moderately different in the psychometric characteristic (i.e., mean of the *b*-parameters) and the percentages of items per content category through Stocking and Swanson (1996)'s weighted deviation model (WDM). One set of five nonparallel item pools had 240 items and the other set of five pools had 480 items. CAT simulations were performed based on the assembled item pools. The MLE procedure was used for scoring, and the test length was 30 items. Item exposure control was imposed. Stocking and Swanson (1993)'s WDM was used for the item selection to balance the content categories.

Overall, the results indicated that the precision of ability estimates was not affected by the moderate change of the psychometric characteristic across item pools. However, the nonparallel pools in terms of proportion of items per content category had an effect on the number of items administered in each content category.

Effects of Nonequivalence of Item Pools on Ability Estimates in CAT

Introduction

Multiple item pools are often needed in a high-stakes computerized adaptive testing (CAT) program as items in an item pool become obsolete or overexposed as time goes on (Stocking, 1994; Way, Steffen, & Anderson,1998). When there are multiple item pools formed from a large item inventory, or formed by adding and deleting items from an existing pool, it is important for the adaptive tests to be parallel to each other regardless of which pools the adaptive tests are from. Also, the resulting reported scale scores should be as similar as possible in the psychometric properties over different pools and be comparable to each other (Stocking, 1994; Wang & Kolen, 1997; Way, Steffen, & Anderson, 1998).

There may be a need to keep multiple pools as parallel as possible in order to maintain the comparability of the reported scale scores derived from CAT tests based on those pools (Wang & Kolen, 1997). Stocking (1994, p. 30) pointed out that some differences in characteristics of item pools (e.g. number of items) can affect the parallelism of the adaptive tests constructed from multiple pools. The equivalence of item pools may be preferable to keep score comparability when the number of items in each pool is small, leading to a lack of sufficient information available in each pool. Also, from an examinees' perspective, parallel pools seem to be more fair when multiple pools are rotated.

Content balance and psychometric characteristics of items are important considerations in constructing multiple pools. Stocking and Swanson (1996) implied that multiple pools can be constructed to be parallel in terms of content and psychometric properties. So, parallel or equivalent pools in this study refer to multiple pools that are constructed to be equivalent in terms of number of items per content and psychometric properties (particularly observed mean of the *b*-parameters).

In practice, however, it is not easy to construct and maintain multiple parallel pools. Way, Steffen, and Anderson (1998, p. 7-9) enumerated several factors impacting the maintenance of the item vat or the universe of items: the shortage of items in the vat, interactions of item characteristics and docking rules (i.e., rules for item retirement), unequal performance of items for the duration of usable life, items becoming obsolete, detection of flawed items, and test disclosure requirements. These factors could also make it difficult to create multiple pools that are parallel. Automated test assembly algorithms such as Stocking and Swanson (1996)'s weighted deviation model (WDM) and the linear programming procedure by van der Linden and Luecht (1998) could be applied to assemble multiple item pools comparable to each other. However, these algorithms have some problems in assembling multiple parallel pools such as frequent infeasible solutions using the linear programming procedure or no guarantee of meeting all constraints using the WDM (Wang, Fan, Yi, Ban, & Zhu, 2000).

Parallel item pools are preferable in some situations to obtain score comparability but, in practice, it is not easy to construct and maintain parallel multiple pools over years of continual testing. So, it is valuable to investigate how nonparallel pools affect score comparability.

The purpose of this paper was to investigate effects of nonparallel item pools on examinees' ability estimates and on the number of administered items in each content category. Because ability estimates obtained in CAT, typically, are transformed to the reported scale score, the examination of ability estimates across item pools provides evidence for score comparability. Also, the number of administered items in each content category is an important consideration for score comparability.

Methods

Data

This study used 39 60-item ACT Mathematics test forms (ACT, 1997). These test forms were administered in paper and pencil (P&P) mode over 5 years. The computer program BILOG (Mislevy & Bock, 1990) was used to estimate item parameters for all items assuming a three-parameter logistic IRT model. The test forms within the same year were administered to randomly equivalent groups of about 2,300 examinees.

The item parameter estimates of the forms administered within the same year were on the same scale but the item parameter estimates of the forms administered in different years were not on the same scale. The anchor form design was used to link forms over different years, and the Stocking and Lord (1983) procedure was used to put item parameter estimates for the forms administered in different years on the same scale. A total of 2,340 items that were on the same scale constituted the final item vat.

Construction of Nonparallel Item Pools

From this item vat, two sets of five item pools were assembled using the WDM: one set of five pools with 240 items in each pool, and the other set of five pools with 480 items in each pool. In constructing the five pools of each set, six content categories (PA, EA, IA, CG, PG, and TG; see ACT, 1997) and the mean of the *b*-parameters were used as constraints. Table 1 shows the means and standard deviations for the item parameters of the five item pools constructed with 240 items and 480 items. Table 2 shows the percentages of items in each content category. Two item pools in each set, called here Basepool1 and Basepool2, were used as baseline item pools. Two baseline pools were assembled to be as similar as possible in terms of the mean of the *b*-parameters and the number of items in each content category. The item pool Contpool was constructed to be similar to the two baseline pools for the mean of the *b*-parameters, but different in the percentage of items in each content category. The item pool Statpool was created to be the same as the two baseline pools in the percentage of items in each content category, but different in the mean of the *b*-parameters. Finally, we constructed the pool Stat&Contpool to be different from the baseline pools in both the mean of the *b*-parameters and the percentage of items in each content. The means of the *b*-parameters and the percentages of item in the content categories were manipulated to be moderately (rather than largely) different for the nonparallel pools to reflect reality.

Table 1. Mean and Standard Deviation for Item Parameters of the Five Item Pools

240-Items Pool	a		b		c	
	Mean	SD	Mean	SD	Mean	SD
Basepool1	1.087	0.349	0.137	1.092	0.194	0.090
Basepool2	1.021	0.297	0.133	1.063	0.184	0.080
Contpool	1.062	0.324	0.166	1.061	0.193	0.087
Statpool	1.024	0.308	-0.079	1.044	0.192	0.083
Stat&Contpool	1.029	0.322	-0.049	1.057	0.192	0.086
480-Items Pool						
Basepool1	1.058	0.336	0.127	1.078	0.187	0.081
Basepool2	1.052	0.336	0.140	1.059	0.189	0.085
Contpool	1.044	0.317	0.120	1.074	0.190	0.085
Statpool	1.030	0.319	-0.057	1.053	0.193	0.086
Stat&Contpool	1.023	0.310	-0.081	1.052	0.191	0.085

The two baseline pools were constructed to be exclusive to each other: no overlapping items were allowed. This was done to examine the similarity of CAT results produced by parallel item pools. The other item pools were overlapping: some items appeared in multiple pools. Overlapping items across item pools reflect a realistic scenario for assembling multiple item pools.

Table 2. Percentages of Items in Each Content Category for 240- and 480-Items Pools

	Content Categories					
Item Pool	PA	EA	IA	CG	PG	TG
Basepool1	23	17	15	15	23	7
Basepool2	23	17	15	15	23	7
Contpool	20	20	20	10	20	10
Statpool	23	17	15	15	23	7
Stat&Contpool	20	20	20	10	20	10

CAT Simulation Procedures

CAT simulations were performed based on the assembled item pools. Fixed-length adaptive tests (30 items) were administered to simulees. Simulees' ability had a rectangular distribution representing 10,500 simulees with 500 simulees having true abilities at each of the 21 equally spaced ability points (in increments of 0.4) beginning with –4.0 and ending with +4.0. The calibrated item parameter estimates were treated as true values. Based on the item parameters and simulated θ s, the probability of an examinee answering an item correctly according to the three-parameter IRT model (i.e., *p*-value) was calculated. This *p*-value then was compared to a randomly generated number (x) from a uniform distribution $U(0, 1)$. If the *p*-value is larger than or equal to the uniform random number, then the simulee got a correct response; otherwise, an incorrect response was obtained for the item. The simulees' abilities were scored using the maximum likelihood estimation (MLE) procedure. For unbounded item responses (i.e., all incorrect or correct responses), the MLE $\hat{\theta}$ was set to -4 or 4, respectively. The CATs began with an item of medium difficulty. The WDM procedure (Stocking & Swanson, 1993) was used for the CAT item selection to balance the content categories. In implementing the WDM procedure, weights were assigned to item information and to the contribution of items in reducing the constraint deviations. Because item information and constraint deviations were not on the same scale, the actual values we gave to these weights did not have definite meaning. We reached these weights by experimenting different values. There were targeted minimum and maximum number of items per content. Simpson and Hatter (1985)'s item exposure control procedure was applied. The upper limit of exposure rate was 0.14 and 0.24 for the 240 items pool and the 480 items pool, respectively.

Criteria

The effects of nonparallel pools on ability estimates were evaluated using bias, standard error (SE), and root mean squared error (RMSE) indices. The three error indices were calculated as following:

$$Bias(\hat{\theta} \mid \theta_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} (\hat{\theta}_i - \theta_k),$$

$$SE(\hat{\theta} \mid \theta_k) = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} (\hat{\theta}_i - \frac{\sum_{i=1}^{n_k} \hat{\theta}_i}{n_k})^2},$$

$$RMSE(\hat{\theta} \mid \theta_k) = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} (\hat{\theta}_i - \theta_k)^2},$$

where θ_k is the true ability at ability level k, $\hat{\theta}_i$ is the estimated ability for simulee i, and n_k represents the number of simulees at ability level k, which is 500. Here, $\text{RMSE}^2 = \text{SE}^2 + \text{Bias}^2$.

The weighted average deviation of administered items in each content category was also computed. The weight was based on a *N*(0, 1) distribution over ability levels (-4, +4). This information shows on average how much the number of administered items deviate from the targeted number of items in each content category.

Results

The empirical results of the effects of nonparallel pools on ability estimates appear in Figures 1 and 2 for the 240- and 480-items pools. Figures 3 and 4 show the average deviation of administered items in each content category for the different item pool sizes.

Bias

The top of Figure 1 shows the conditional bias plots of the five pools for the 240 items pool. These plots indicate that the bias for Contpool tends to be relatively larger at some lower ability levels than that for the other four pool (Basepool1, Basepool2, Stat&Contpool, and Statpool). Compared to the differences between the two base pools, however, this difference does not appear to be significant. At the middle ability levels, the biases were not distinguishable among the five pools. The biases for the pools at the higher ability levels were similar to each other. For the 480 items pool in Figure 2, the differences of conditional biases for the five pools were small across all ability levels.

The conditional bias plots in Figures 1 and 2 shows that the pool differences in terms of the mean of the *b*-parameters and/or the contents may not affect the bias in ability estimation.

Standard Error

In the middle of Figure 1, the conditional standard errors of the five pools are plotted for the 240 items pool. Figure 1 shows that the standard errors for Statpool and Stat&Contpool tend to be smaller than that for Basepool1, Basepool2, and Contpool at lower ability levels. This makes sense because the means of the *b*-parameter of these two pools were somewhat smaller than those of the other pools. At the middle ability levels, the differences of the standard error were negligible among the five item pools. The conditional standard errors for Contpool were larger than that for the other pools only at lower ability levels. At the high ability levels, the differences of the conditional standard errors for the pools appear to be not large compared to the differences between the two base pools.

Figure 2 shows that for the 480 items pool, the standard errors for Statpool and Stat&Contpool were smaller at the low ability levels than that for the other pools. This result is consistent with that of the 240 item pools. At middle ability levels, the standard errors show minor differences among the pools. Stat&Contpool produced larger standard errors, sometimes, at the high ability levels than the other pools did.

The differences of the conditional standard errors for the pools shown in Figures 1 and 2, however, appear to be largely due to randomness when the results were compared to the differences in the conditional standard errors between the two base pools.

RMSE

The bottom of Figure 1 shows the conditional RMSE plots of the five pools for the 240 items pool. Because RMSE is a function of both bias and SE, the results for RMSE are related to those already provided for bias and SE. Statpool and Stat&Contpool produced smaller RMSEs at the low ability levels than that for the other pools. The larger RMSE was produced for Contpool at the low ability levels. At the middle ability levels, there were minor differences in RMSE among the pools. At the high ability levels, the conditional RMSEs appear to be not different among the pools.

For the 480 items pool in Figure 2, the RMSEs for Statpool and Stat&Contpool were smaller at the low ability levels than that for the other pool. However, overall, all the differences of the conditional RMSEs among the pools were minor across all ability levels.

In summary, based on the results obtained from the 240 and 480 items pools, the effect of statistical characteristic difference (i.e., the mean of the *b*-parameters) in item pools on ability estimation appears to be minor under the conditions of this study.

Deviation of Administered Items From Targeted Number of Items

Figure 3 shows the average deviations of administered items in each content category from the targeted number of items for the 240 items pool. Basepool1, Basepool2, and Statpool produced smaller deviations of administered items in the six content categories than Contpool and Cont&Statpool did. The items in EA, IA, and TG for both Contpool and Cont&Statpool were more administered than the other pools, whereas the items in PA, CG, and PG were less administered than the other pools. Both Contpool and Cont&Statpool had more items in the content categories of EA, IA, and TG and less items in the PA, CG, and PG categories than the other pools had.

Deviations of administered items in each content category from the targeted number of items for the 480 items pool are presented in Figure 4. The pattern of the deviations was similar to that of the 240 items pool.

When the proportions of items in the six content categories were changed for the different item pools, the items in the content categories that were increased in proportion appear to be more administered than the items in the other content categories that were decreased in proportion. That is, the greater number of informative items from the content categories that were increased lead to items in these categories being administered more often, which led to relatively large positive deviations in those categories. Consequently, the items in the other content categories were less utilized and large negative deviations were produced.

The deviations (positive or negative) of administered items from the targeted number of items were similar for both the 240 items pool and the 480 items pool, except for Basepool1. Because the upper limit of item exposure rates were set to 0.14 and 0.24 for the 240 items and 480 items pools, respectively, the effect of different item pool sizes on the deviations of administered items was controlled by the different levels of the item exposure rates.

Based on the results obtained from the 240 and 480 items pools, different proportions of items per content category across the pools had an effect on the number of administered items per content category. A practical implication of the results is that when the proportions of items per item content category are changed across different pools, the number of items to be administered per content category should be controlled. Otherwise, many informative items in some content categories will be more often administered, which will result in a large deviation of administered items per content category from the targeted number of items to be administered.

Conclusion and Discussion

The purpose of this study was to evaluate the effect of nonparallel multiple item pools on ability estimation and the number of administered items per content category, where nonparallel item pools were constructed to be moderately different in the psychometric characteristic (i.e., mean of the *b*-parameters) and the percentages of items per content.

Overall, the results indicated that the precision of ability estimates was not affected by the moderate change of the psychometric characteristic across item pools. However, the nonparallel pools in terms of proportion of items per content category had an effect on the number of items administered in each content category. When the proportions of items in some content categories were increased, deviations of the administered items from the targeted number of items in these content categories became also larger. The large deviations of administered items from the targeted items for that pool, however, may be controlled by narrowing the range of the minimum and maximum number of items to be administered in each content category.

It is not uncommon for a high-stakes CAT program to confront moderately nonparallel multiple pools in terms of the psychometric characteristic (i.e., mean of the *b*-parameters) and the percentages of items per content category. The results of this study appear to be somewhat encouraging for CAT developers, because the precision of ability estimation did not appear to be affected much by the nonequivalent item pools. Also, the number of administered items in each content category may be controlled by a sophisticated item selection algorithm.

It is emphasized that the results reported here should be interpreted with caution due to the use of the limited definition of the nonparallel item pools. This study defined nonparallel pools in terms of only mean of the *b*-parameters and the proportion of items in each content category, and used this definition to assemble the multiple pool. The actual multiple item pools for a CAT may be much different from the pools used in this study. For example, multiple item pools could be different from each other in terms of mean of the *a*-parameters, overall item information function, exposure rate, difficulty and discrimination of items within each content category, and so on. Nonparallel pools in terms of these factors may or may not have an effect on ability estimates and the number of administered items per content category.

It should also be cautioned that the results reported in this paper may be closely related particularly to the CAT item selection method (i.e., the WDM method) and to the weights assigned to information in implementing this method. Changes in the weights or in the item selection method itself might change some of the results.

The nature and characteristics of item pools for CAT would be very different among testing programs and within a testing program over years of continual testing. It is necessary for test developers to conduct their own simulation analyses over years to assure score comparability across nonparallel multiple pools and across modified versions of an item pool.

Reference

Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model* (2nd ed.). Mooresville, IN: Scientific Software.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 17th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Stocking, M. L., & Swanson, L. (1996). *Optimal design of item pools for computerized adaptive testing* (Research Report 96-34). Princeton, NJ: Educational Testing Service.

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

van der Linden, W. J., & Luecht, R. R. (1998). Observed-score equating as a test assembly problem. *Psychometrika, 54*, 237-247.

Wang, T., & Kolen, M. J. (1997). *Evaluating comparability in computerized adaptive testing: A theoretical framework with an example.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wang, T., Fan, M., Yi, Q., Ban, J.-C., & Zhu, D. (2000). *Assembling parallel item pools for computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Way, W. D., Steffen, M., & Anderson, G. S. (September, 1998). *Developing, maintaining, and renewing the item inventory to support computer-based testing.* Paper presented at the colloquium, Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

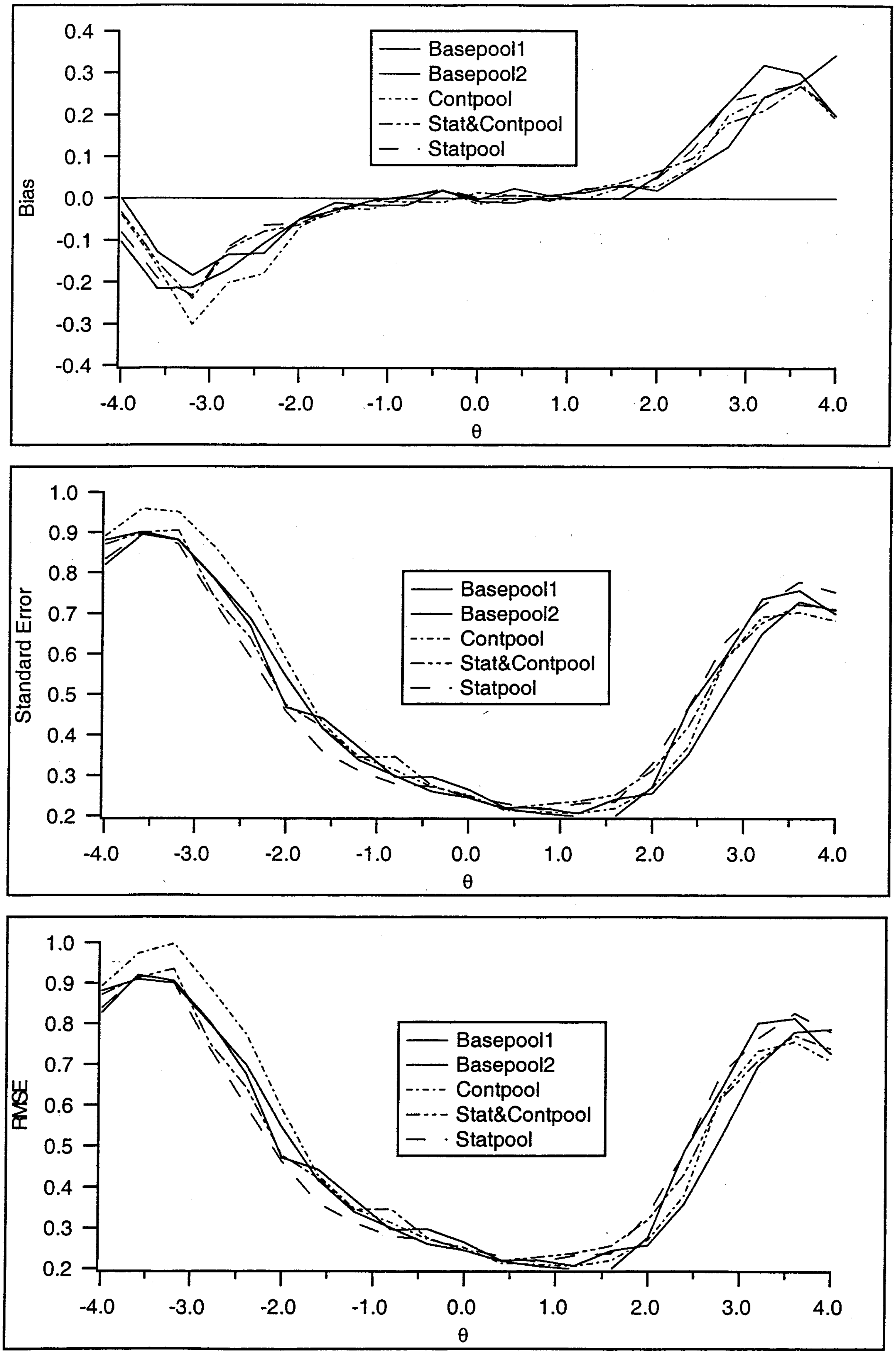

Figure 1. Bias, SE, and RMSE for the 240-items pool

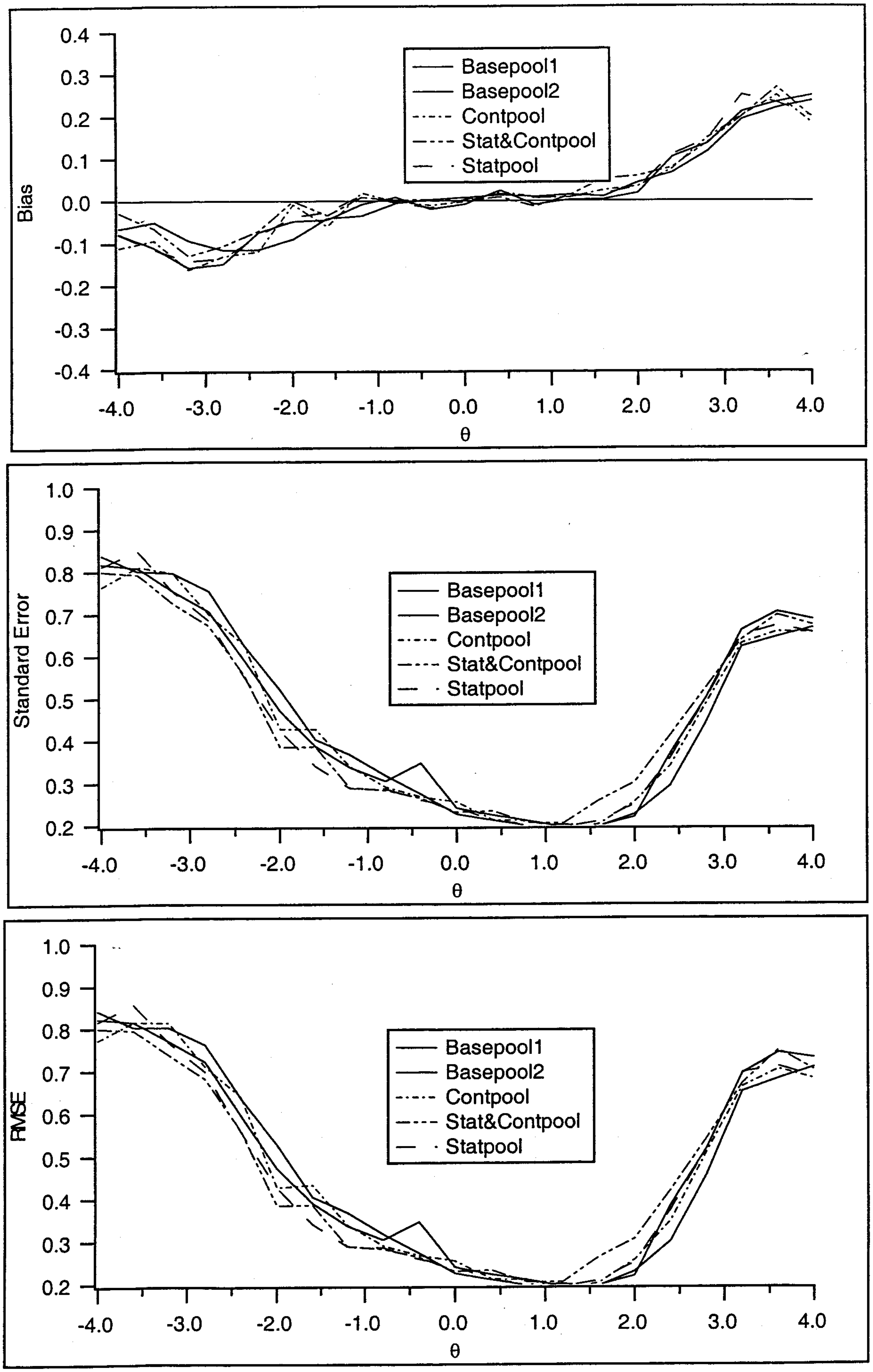

Figure 2. Bias, SE, and RMSE for the 480-items pool

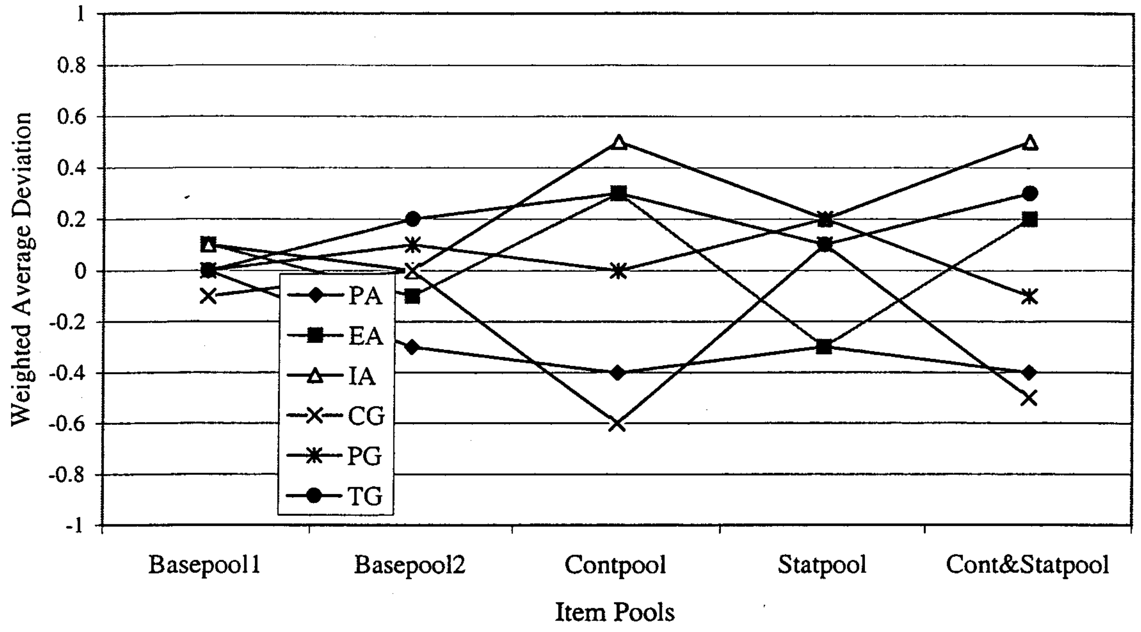

Figure 3. Average Deviation of Administered Items on Each Category Using the 240 Items Pool

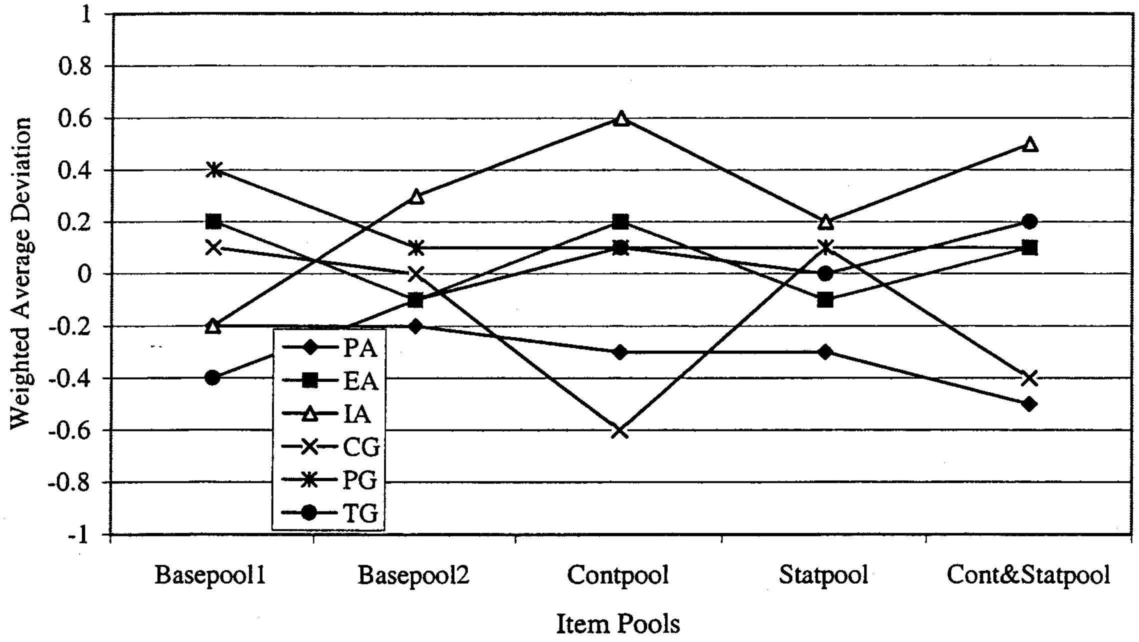

Figure 4. Average Deviation of Administered Items on Each Category Using the 480 Items Pool