DOCUMENT RESUME

ED 476 923 TM 034 974

AUTHOR	Pommerich, Mary; Harris, Deborah J.
TITLE	Context Effects in Pretesting: Impact on Item Statistics and Examinee Scores.
PUB DATE	2003-04-00
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Context Effect; *Pretesting; Simulation; *Statistical Bias; Test Construction; *Test Items

ABSTRACT

In this study, the effect of appended pretesting was evaluated with regard to item statistics and examinee scores for groups of items that were pretested as part of a large-scale operational testing program. In appended pretesting, items are administered in a separately timed section at the end of an operational test battery. Two evaluations were conducted: one using a pretest unit consisting of a reading passage and the other using a pretest unit consisting of mathematics items, most of which were discrete. Sample sizes were: (1) 634 for the original pretesting group; (2) 294,637 for the operational group; (3) 1,007 for the re-pretest as operational group; and (4) 1,021 for the re-pretest as original group. In the ideal case, if items are pretested in exactly the context in which they will appear operationally, there should be no context effects. Results from the simulations in this paper showed some small-to-moderate negative effects on scores when misspecified parameters were used for item response theory scoring. Larger negative effects occurred at score points where fewer examinees score. Score bias could be smaller or larger under different conditions from those observed in this study. Results from the study also support the idea that preequating should only be conducted under very carefully controlled situations. (Contains 12 tables and 6 references.) (SLD)

Context Effects in Pretesting: Impact on Item Statistics and Examinee Scores

Mary Pommerich
Defense Manpower Data Center

Deborah J. Harris
ACT, Inc.

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April 2003.

The views expressed are those of the authors and not necessarily those of the Department of Defense, or the United States Government.

The authors thank Jay Happel for his patience and time in supplying the data for this study, and Tim Burden for his assistance in designing the study.

Context Effects in Pretesting: Impact on Item Statistics and Examinee Scores

An essential part of the test development process is the pretesting of items. Pretesting allows us to gather information about an item that is being considered for placement on an operational test form or in a CAT item pool. The information gathered can be used to make decisions about items to include or exclude from operational forms, and to assist with building forms or item pools to desired levels of difficulty.

Pretesting may be conducted as part of an operational test administration, or in a special study that is not affiliated with an operational test administration. This paper considers only the case where pretesting is conducted as part of an operational test administration, and focuses on two types of pretesting: appended and embedded pretesting. In appended pretesting, items are administered in a separately-timed section at the end of an operational test battery, while in embedded pretesting, items are administered within an operational test. Appended pretesting ensures that an examinee's operational scores are not influenced by the pretest items, but it is possible that astute examinees can identify that those items will not contribute toward their test scores, and give less effort than they would on the operational items. Motivation may be a consistent problem associated with appended pretesting. Embedded pretesting makes it more difficult to identify pretest items so that examinees may give the same level of effort as on operational items. However, the inclusion of pretest items in an operational section can possibly influence examinee performance on operational items, thereby affecting operational scores. Appended pretesting is a more conservative approach than embedded pretesting in that it will not affect an examinee's operational score, but it is a less desirable practice from a psychometric perspective, because of the potential that examinees behave differently than they would on an operational test.

Pretest items are also typically administered in a somewhat different context than they are administered operationally. A group of pretest items are administered together in a pretest "unit." When operational forms or item pools are built, a pretest unit is typically not placed intact on the operational form. Items on an operational form may come from multiple pretest units. Items on an operational form may appear with a completely different group of items than they were pretested with. For passage-based items, items may be dropped between pretest and operational administrations, and the items may appear in different orders across the pretest and operational administrations. In addition to potential administration context effects, there may be context effects due to a lack of motivation for appended pretesting situations. All of this creates the possibility that inferences about items that are made based on a pretest administration may not hold in a subsequent operational administration.

Study Design

In this study, the effect of appended pretesting was evaluated with regard to item statistics and examinee scores, for groups of items that were pretested as part of a large-scale operational testing program. A group of items that are pretested together is referred to as a "pretest unit." Two evaluations were conducted: one using a pretest unit consisting of a reading passage, the other using a pretest unit consisting of math items, most of which were discrete. The math unit did contain one "item set" – three items associated with the same prompt. Where pretesting was conducted, appended pretesting was used.

In the first evaluation, a reading passage that had previously been pretested and then administered as part of an operational test form was "re-pretested" in two different ways: using the same item order as when the passage was originally pretested, and using the same item order as when the passage was operationally administered. The content of the reading passage and the items themselves remained the same, but the item order differed. The order of response options was also changed for a handful of items. The repretesting under the original pretest and operational conditions allowed the comparison of item performance and item statistics across the different item orders, under controlled conditions.

In the second evaluation, a completely new math pretest unit was administered, using two different item orders. This unit had never been pretested or operationally administered before. The administration of the scrambled math units allows further comparison of order effects for a primarily discrete item test, again under controlled conditions.

Administration History for the Reading Passage

Table 1 summarizes the administration history for the reading passage, with regard to the item and response foil order used across the different administrations. The reading passage was originally pretested in 1996 (labeled "Original Pretest"). It consisted of a passage and 15 items. Each item contained four response foils. The order in which the items appeared in this administration is used as the reference order for all subsequent administrations. Item orders for the operational administration and the two re-pretest administrations are given with reference to this original order. The foil order for the four response options is also given in parentheses. Foil orders for the operational administration and the two re-pretest administrations are also given with reference to the original pretest foil order.

After the original pretesting, the reading passage was placed onto an operational test form that was administered operationally in 2000 (labeled "Operational"). The operational form consisted of four reading passages containing 10 items each. The reading passage of interest appeared as the third passage of the operational form. Five of the items that were originally pretested with this passage were not included with the passage when it was administered operationally. Items 1, 5, 8, 13, and 15 from the original pretest unit did not appear with this passage on the operational form. In addition, the order in which the items appeared changed from the original pretest to the operational administration. For example, Item 2 from the original pretest administration appeared as Item 6 of the passage in the operational administration (which in turn was Item 26 on the entire test). Item 1 from the operational administration (which in turn was Item 21 on the entire test) appeared as Item 11 in the original pretest administration. The content of the passage and items that were administered were unchanged across the original pretest and operational administrations. In addition, the response foil orders were unchanged across the original pretest and operational administrations.

Table 1. Item Order (and Foil Order) for the Reading Passage Across Original Pretest, Operational, and Re-Pretest Administrations.

Original Pretest (1996)	Operational (2000)	Re-Pretest As Original Pretest (2001)	Re-Pretest As Operational (2001)
1 (ABCD)	Not Administered	1 (ABCD)	11 (DCBA)
2 (ABCD)	6 (ABCD)	2 (ABCD)	6 (ABCD)
3 (ABCD)	5 (ABCD)	3 (ABCD)	5 (ABCD)
4 (ABCD)	7 (ABCD)	4 (ABCD)	7 (ABCD)
5 (ABCD)	Not Administered	5 (ABCD)	12 (CDAB)
6 (ABCD)	9 (ABCD)	6 (ABCD)	9 (ABCD)
7 (ABCD)	8 (ABCD)	7 (ABCD)	8 (ABCD)
8 (ABCD)	Not Administered	8 (ABCD)	14 (CDAB)
9 (ABCD)	10 (ABCD)	9 (ABCD)	10 (ABCD)
10 (ABCD)	2 (ABCD)	10 (ABCD)	2 (ABCD)
11 (ABCD)	1 (ABCD)	11 (ABCD)	1 (ABCD)
12 (ABCD)	4 (ABCD)	12 (ABCD)	4 (ABCD)
13 (ABCD)	Not Administered	13 (ABCD)	13 (DCBA)
14 (ABCD)	3 (ABCD)	14 (ABCD)	3 (ABCD)
15 (ABCD)	Not Administered	15 (ABCD)	15 (ABCD)

In 2001, the original reading passage was re-pretested in two ways. First, the passage was re-pretested under the same conditions as it was originally pretested (labeled "Re-Pretest As Original Pretest"). Namely, all 15 items originally pretested with the passage were re-pretested in the same order as they appeared in the original pretest unit. In addition, the same foil orders were used for all items across the re-pretest and original pretest administrations. Second, the passage was re-pretested to replicate the operational administration (labeled "Re-Pretest As Operational"). The examinee groups taking the passage under the Re-Pretest As Original Pretest and Re-Pretest As Operational administrations were randomly equivalent (the pretest passages were spiraled across groups). The only difference across these two administrations was item order and response foil order for four items.

Under the Re-Pretest As Operational administration, there were some necessary changes to account for the different number of items across operational and pretest administrations. The 10 items that appeared on the operational form were re-pretested in the same order as they appeared on the operational form (with the same foil order). For example, Item 1 from the operational administration was Item 1 in the Re-Pretest As Operational administration. (This same item appeared as Item 11 in the Original Pretest and Re-Pretest As Original administrations.) The five items that were originally pretested but did not appear on the operational form were appended in positions 11-15 of the Re-Pretest As Operational administration. The foil order for Items 11-14 of the Re-Pretest As Operational administration was modified from that of the Original Pretest administration. Foil orders were reversed for Items 11 and 13 of the Re-Pretest As Operational administration (i.e., response options appearing as ABCD in the Original Pretest administration appeared as DCBA in the Re-Pretest as Operational administration). Foil orders were inverted for Items 11 and 14 of the Re-Pretest As Operational administration (i.e., response options appearing as ABCD in the Original Pretest administration appeared as CDAB in the Re-Pretest as Operational administration). The foil order for Item 15 of the Re-Pretest As Operational administration was unchanged, so that Item 15 was identical across the Original Pretest and Re-Pretest As Operational administrations (which in turn was identical to Item 15 in the Re-Pretest As Original administration).

Administration History for the Math Unit

Table 2 summarizes the administration history for the Math unit, with regard to the item order used in the two pretest administrations. The unit contained 16 items from the content areas of Pre-Algebra, Elementary Algebra, Plane Geometry, Coordinate Geometry, and Trigonometry. The unit was pretested for the first time in 2001, using two different item orders. The order in which the items appeared in the administration labeled "Pretest Order 1" is used as the reference order for the pretest administration labeled "Pretest Order 2." For example, Item 1 in the Pretest Order 1 administration was Item 13 in the Pretest Order 2 administration. The item content was identical across the two pretest administrations, as was the foil order. The examinee groups taking the unit under the Pretest Order 1 and Pretest Order 2 administrations were randomly equivalent (the pretest units were spiraled across groups).

Table 2. Item Order and Item Type for the Math Unit Across the Pretest Administrations.

Pretest Order 1 (2001)	Pretest Order 2 (2001)	Item Type
1	13	Set
2	14	Set
3	15	Set
4	12	Discrete
5	5	Discrete
6	11	Discrete
7	9	Discrete
8	8	Discrete
9	7	Discrete
10	2	Discrete
11	3	Discrete
12	6	Discrete
13	4	Discrete
14	10	Discrete
15	1	Discrete
16	16	Discrete

Results for Reading

Initial Data Screening

Data from each of the three administration dates (1996, 2000, and 2001) were screened to eliminate examinees that were not in grades 10-12, which is the typical testing population for the operational testing program. Examinees outside of this grade range could be affected differently by context effects than the typical testing population. Resulting sample sizes were N = 634 for the Original Pretest group, N = 294,637 for the Operational group, N = 1007 for the Re-Pretest As Operational group, and N = 1021 for the Re-Pretest As Original Pretest group. A random sample of N = 1007 was created for the Operational group, for purposes of comparison with the Re-Pretest As Operational group. This smaller random sample was used for all analyses of the Operational group.

The level of motivation for the re-pretest groups on the pretest unit was also evaluated. The distribution of scale scores for the two groups on the operational reading test was approximately normally distributed. The distribution of raw scores on the re-pretest unit was negatively skewed across the two groups. This suggests both that the pretest unit may have been easier overall than the operational test as a whole, and that there may have been some low-motivated examinees, pulling the mean raw score down. Some of the examinees may have performed differently on the pretest unit than they did on the operational reading test.

The random equivalence of the Re-Pretested As Operational and the Re-Pretested As Original Pretest groups was evaluated to ensure that performance could be compared across the two re-pretest administrations. A t-test of the operational reading scores showed no significant differences across the two re-pretest groups. A t-test of the operational total battery scores showed no significant differences across the two re-pretest groups. In addition, χ^2 tests of independence between re-pretest group and gender, ethnicity, and grade were not significant, suggesting no relationship between re-pretest group and gender, ethnicity, or grade. Thus, the re-pretested groups appeared to be randomly equivalent.

Evaluation of Item P-Values

Re-Pretest As Original Pretest Versus Re-Pretest As Operational

Item p-values for each re-pretest group are given in Table 3. Items from the Re-Pretested As Operational condition were reordered to match the order in the Re-Pretested As Original Pretest condition. Item foil order for Items 11-14 from the Re-Pretest As Operational condition was also reordered to match the order in the Re-Pretest As Original Pretest condition. The p-value differences (± 2 standard errors) are plotted in Figure 1. Error bands that do not surround the zero line suggest significantly different p-values across the two re-pretest administration groups. Items 4, 5, and 11 all showed p-value differences that significantly favored the Re-Pretest As Original Pretest Group. The item order appeared to have some effect on item difficulty for some items, but not others. A t-test of the raw score on the pretest unit showed no significant difference across the two re-pretest groups.

χ^2 tests of independence between re-pretest group and item score, testing whether there was a relationship between re-pretest group and item score, were also significant ($p < .01$) for Items 4, 5, and 11. χ^2 tests of independence between re-pretest group and response, testing whether there was a relationship between re-pretest group and the response option chosen, were significant ($p < .05$) for Items 2 and 14 (in addition to being significant for Items 4, 5, and 11). The tests of group by response suggest that examinees responded at different rates to the different foils across the re-pretest groups on these five items. On Items 2 and 14, the examinees appeared to have responded at different rates among the incorrect foils. All of the items appeared in different positions across the two administrations. In addition, Item 5 had different foil orders across the two administration conditions.

Original Pretest Versus Re-Pretest As Original Pretest
The comparison of item p-values across the Original Pretest and Re-Pretest As Original Pretest groups is interesting, but somewhat difficult to interpret because these adminstrations were conducted at different points in time to groups that were not randomly equivalent. The Original Pretest and Re-Pretest As Original Pretest administrations differed only in terms of the number of years between testing (5). The same pretest unit was administered in each of these cases, at the same time of year, under the same administration conditions. χ^2 tests of independence between administration group and gender, ethnicity, and grade were not significant, suggesting no relationship between administration group and gender, ethnicity, or grade. A t-test of the operational reading score showed no significant differences across the two administration groups. A t-test of the operational total battery score did show a significant difference across the two re-pretest groups ($p < .01$). A t-test of the raw score on the pretest unit showed no significant difference across the two administration groups.

Table 3. Item P-Values for the Two Re-Pretest Administrations.

Re-Pretested As Original Pretest		Re-Pretested As Operational	
Item Position	P-Value	Item Position	P-Value
1	0.87	11	0.89
2	0.66	6	0.64
3	0.68	5	0.66
4	0.62	7	0.54
5	0.81	12	0.76
6	0.52	9	0.52
7	0.59	8	0.62
8	0.88	14	0.87
9	0.53	10	0.56
10	0.75	2	0.72
11	0.82	1	0.78
12	0.77	4	0.77
13	0.87	13	0.87
14	0.76	3	0.74
15	0.83	15	0.81

The p-value differences (± 2 standard errors) are plotted in Figure 2. Items 5 and 10 showed p-value differences that significantly favored the Original Pretest group. There is no clear explanation for why these items significantly favored the Original Pretest group over the Re-

Pretest As Original Pretest group. Score trends with the SAT and ACT suggest that examinee populations are growing more able over time, which, if true, should favor the Re-Pretest As Original Pretest group. However, P-value differences could be affected if examinee populations become less motivated over time. There is no evidence to suggest that this is the case, but it is possible that the longer a pretesting practice is used, the more likely examinees are to become aware of its use, especially if coaching schools are aware of the practice. Further comparisons of the Original Pretest and Re-Pretest As Original Pretest groups were not conducted because of the difficulty in interpreting results.

Operational Versus Re-Pretest As Operational

The comparison of item p-values across the Operational and Re-Pretest As Operational groups is even more difficult to interpret because there are many differences across the two administrations. First and foremost, the Operational group took the unit under an operational administration, while the Re-Pretest As Operational group took the unit under a pretest administration. Motivation (or lack of motivation) could be a big factor in influencing performance. In addition, differing numbers of items were administered to each group (10 versus 15), different time constraints were used (35 minutes for 40 total items versus 20 minutes for 15 total items), and the unit was administered at different times of the year in different years. The p-value differences (± 2 standard errors) are plotted in Figure 3. All items but 2 and 6 showed significant differences, some items favoring the Operational group, others favoring the Re-Pretest As Operational groups. There is no clear explanation for the differences observed. Further comparisons of the Operational and Re-Pretest As Operational groups were not conducted because of the difficulty in interpreting results.

Evaluation of IRT Parameters

The p-value differences suggest that items differed in difficulty across the different administration conditions. IRT DIF methods were used to evaluate differences in IRT difficulty, discrimination, and guessing parameters across the Re-Pretested As Original Pretest and Re-Pretested As Operational groups. Likelihood-ratio chi-square tests (Thissen, Steinberg, and Wainer, 1993) were used to test the equality of IRT parameters across the two groups. The program IRTLRDIF v.2.0b (Thissen, 2001a, Thissen, 2001b) was used to conduct the likelihood ratio tests. In this program, items that display significant differences in item characteristic curves over the groups are evaluated for DIF at the item parameter level. Items 1, 4, 5, 7, 8, 9, and 10 all displayed significant differences ($p < .05$) in at least one parameter over the two re-pretest groups. The a, b, and c parameters all displayed significant differences across the re-pretest groups, for at least one item.

Evaluation of Impact on Scores

The results of the p-value analyses and the IRT DIF analyses for the two re-pretest groups suggest that administering the items in different orders can affect item difficulty, item discrimination, and guessing for individual items. Because the groups were randomly equivalent, and the pretest units were administered under the same, controlled conditions, the difference in performance can be attributed to the difference in item order across the two administrations. The findings suggest that the context (in this case, order) in which an item appears may affect its performance. In the case of the pretest administrations versus the operational administration, the context differed both in terms of item order and type of administration, which could have contributed to even larger performance differences. We cannot draw that conclusion directly because the pretest and operational administrations were not administered under controlled conditions that would allow us to isolate the causes of performance differences. But there is some evidence to suggest that analyses of data from one administration type versus the other could result in different item statistics and different impressions about the characteristics of the items.

Administering items in different contexts may create differences in item statistics across the administration groups, but even significant differences in item statistics may not be practically significant if examinees receive the same score, regardless of the order in which they take the items. A simulation study was conducted to compare the effect of using item calibrations from a "pretest administration" versus item calibrations from an "operational administration" to obtain IRT scores, for items that were administered operationally. The simulation was based on real data from the Re-Pretest As Original and Re-Pretest As Operational administrations. Because the two re-pretest groups took the pretest units under the same administration condition (appended pretesting), differences in motivation due to administration type (pretest vs. operational) could not be simulated. Thus, the simulated effects due to "pretest administration" versus "operational administration" consisted only of order effects, corresponding to the order effects observed across the two re-pretest conditions. The order effects are believed to be a realistic representation of differences that might occur across pretest and operational administrations, as the orders used in the two re-pretest conditions matched those used in the Original Pretest and Operational administrations for the real data (with the exception of the 5 items appended to the end of the 10 operational items for the Re-Pretest As Operational condition).

The item order from the Re-Pretest As Operational administration was treated as the order in which items would be administered operationally. All item responses were therefore generated using the item parameters from calibrations of the real data from the Re-Pretest As Operational administration. These parameters were considered to be the "true" parameters, because our interest was in scores examinees receive on the operational administration. Item calibrations were also conducted using real data from the Re-Pretest As Original Pretest administration. These parameters were considered to be "mis-specified" to some degree, because the calibrations were based on items administered in a different context (i.e., differing item order) from the operational administration. IRT scores were computed using both the true and mis-specified parameters. For an operational administration, the ideal case from a psychometric perspective would be that examinees respond to items in the same context as which their IRT scores are derived (i.e., data from the operational administration is also used to obtain item parameters used in scoring). In a less-than-ideal case, examinees would respond to items in a different context from which their IRT scores are derived (i.e., data from a pretest administration is used to obtain item parameters used in scoring data from an operational administration). A comparison of scores based on the true and mis-specified parameters will show the effect of using the ideal (true) and less-than-ideal (mis-specified) parameters when scoring responses from an operational administration.

The simulation study was conducted to mimic the degree of parameter differences observed across the Re-Pretest As Original Pretest and Re-Pretest As Operational administrations. Bilog-MG for Windows (Zimowski, Muraki, Mislevy, and Bock, 2003) was used to conduct separate calibrations using item responses from the Re-Pretest As Operational and Re-Pretest As Original Pretest administrations. Items in the Re-Pretest As Operational administration were reordered to match the item order from the Re-Pretest As Original Pretest administration. Foil orders were also reordered to match the foil order from the Re-Pretest As Original Pretest administration. The item parameters from the Re-Pretest As Original Pretest calibration were rescaled (i.e., linearly transformed) to be placed on the scale of the parameters from the Re-Pretest As Operational calibration. Rescaling had very little effect on the item parameters, because the groups were randomly equivalent.

Two thousand examinees were simulated at each of 13 equally-spaced ability points between ±3.0. In the comparison labeled "True/True," examinees responded according to the true parameters (calibrated from the Re-Pretest As Operational administration) and were scored using the same parameters. In the comparison labeled "True/Mis," examinees responded according to the true parameters (calibrated from the Re-Pretest As Operational administration) and were scored using the mis-specified parameters (calibrated from the Re-Pretest As Original Pretest administration). Maximum likelihood ability estimates (MLE) were computed, along with expected a posteriori (EAP) and Baye's modal estimates (BME). Bias was computed as the difference between the examinee's true and estimated ability. Note that in simulating the test, the parameters for the 15 items were copied to create a test that was four times the length of the pretest unit (60 items). This was done to more clearly demonstrate the effects of the different item parameters on scoring. Results based on the original unit (15 items) and the cloned units (60 items) showed the same patterns and trends, only the differences were more apparent and easier to see with the 60 item test. All simulation results are presented for the cloned units. Note also that only one replication was performed of the simulation. Reporting results over multiple replications could give smoother results than reported here, and provide some indication of the sampling error associated with the simulation results.

Figure 4 shows total test information by calibration group for the two re-pretest units (15 items). The test information function provides an upper bound to the information that can be obtained by any method of scoring the test (Lord, 1980). The information is also plotted for not-rescaled and rescaled parameters for the Re-Pretest As Original Pretest condition. The pretest unit shows quite a bit more information for the Re-Pretest As Operational parameters for abilities between $\theta = -1.8$ and $\theta = 1.4$, and slightly more information for the Re-Pretest As Original Pretest parameters for abilities less than $\theta = -1.8$. At moderate ability levels, where most examinees fall, the Re-Pretest As Operational parameters showed substantially more information than the Re-Pretest As Original Pretest parameters. The similarity of information for the not-rescaled and rescaled parameters for the Re-Pretest As Original Pretest condition demonstrated that the difference in information observed across the Re-Pretest As Operational and Re-Pretest As Original Pretest parameters was not due to the rescaling of the Re-Pretest As Original Pretest parameters. Why one item order versus the other should show such a clear difference in information is uncertain. In operational administrations of the reading passages, items are generally ordered by difficulty, from easier to harder, so that could be a factor.

Figures 5 and 6 present the average absolute bias at each true ability point for the BME and EAP ability estimates, respectively. For the two Bayesian ability estimates, the trends in bias across the True/True and True/Mis conditions follow the trends in test information observed in Figure 4. Namely, where information is higher for the Re-Pretest As Operational administration (at abilities roughly between $\theta = -1.8$ and $\theta = 1.4$), the average absolute bias is lower for the True/True condition (where the Re-Pretest As Operational parameters were used to generate the item responses and score them). Where information is higher for the Re-Pretest As Original Pretest administration (at abilities roughly below $\theta = -1.8$), the average absolute bias is lower for the True/Mis condition (where the Re-Pretest As Operational parameters were used to generate the item responses, and the Re-Pretest As Original Pretest parameters were used to score them). Overall, the difference in bias is pretty minimal, except at $\theta \leq -2.5$, which represent abilities where very few examinees fall. But it is interesting to note that the trend in bias across the two simulation conditions seems so closely tied to the relationship in test information for the two calibration samples.

Figure 7 presents the average absolute bias at each true ability point for the MLE ability estimates. The MLE estimates show more bias for the True/Mis condition at all theta points < 1.5. The MLE estimates also show more bias than the EAP and BME estimates for the True/True condition at both tails of the ability scale. This is likely due to the arbitrary assignment of an MLE estimate of ±5 for examinees with all correct or all incorrect response patterns. Figure 8 presents the average absolute bias for the MLE ability estimates, excluding examinees with scores of ±5. The magnitude of the average absolute bias is now similar to the magnitude of the bias for the BME and EAP estimates. The estimates still show more bias for the True/Mis condition than for the True/True condition, at almost all theta points. The trend in bias across the two simulation conditions does not appear to be as closely tied to the relationship in test information for the MLE estimates as it does for the Bayesian estimates.

Results for Math

Evaluating the results for the math pretest unit is a little different from evaluating the results for the reading pretest unit, because there is no inherent "right" or "wrong" order for the group of items that were tested. For reading, the fact that the items were administered operationally creates a "right" order for evaluation, namely, the operational order. Because the math pretest unit was pretested for the first time in this study, there is no operational order associated with those items, and the "true" calibration condition is unknown. Truth would best be determined by the context in which the items would be administered operationally. Because we do not have that information for the math pretest unit, comparisons of the different item orders only tells us whether or not the items behave differently and the consequences of using the different parameters, with no basis for judging which behavior is better than the other.

Evaluation of Item Statistics

The data for math were screened in the manner described for reading. The two groups were determined to be randomly equivalent so that performance could be compared across the two pretest administrations.

P-value differences (±2 standard errors) across the two item orders are plotted in Figure 9. The order of items for group Pretest Order 1 is used as the reference order, and items from Pretest Order 2 were reordered to match the order of items from Pretest Order 1. Items 3 and 13 were significantly easier for the Pretest Order 1 group than the Pretest Order 2 group.

IRT DIF analyses of the items showed that Items 3, 6, 7, 8, 9, 13, and 15 all displayed significant differences ($p < .05$) in at least one parameter over the two pretest groups. The a, b, and c parameters all displayed significant differences across the pretest groups, for at least one item. Item 1 was a very difficult item (only 11% of examinees answered this item correctly in both pretest groups), and did not converge in the IRT calibrations. Thus, Item 1 was excluded from the evaluation of the effect of the parameter differences on scores.

Evaluation of Impact on Scores

The results of the p-value analyses and the IRT DIF analyses for the two pretest groups suggest that administering the items in different orders can affect item difficulty, item discrimination, and guessing for individual items. As was the case with the reading re-pretest groups, the difference in performance can be attributed to the difference in item order across the two administrations.

A simulation study was conducted in the manner of the simulation for reading, to compare the effect of using item calibrations from the two orders to obtain IRT scores, where one of the orders was designated to be the "true" order. Again, this is a less clear comparison than for reading, because there is no inherent true order for these items. The item order from the Pretest Order 1 group was treated as the order in which items would be administered operationally, and was labeled the true item order. All item responses were generated using the item parameters from calibrations of the real data from the Pretest Order 1 administration. These parameters were treated as the "true" parameters. Item calibrations were also conducted using real data from the Pretest Order 2 administration. These parameters were treated as the "mis-specified" parameters, and are considered to be mis-specified because the calibrations were based on items administered in a different order from the true item order. IRT scores were computed using both the true and mis-specified parameters. Comparisons of scores based on the true and mis-specified parameters will show the effect of using true versus mis-specified parameters to score, when examinees are responding according to the true parameters.

As was done with the reading simulation, the item parameters were cloned four times to create a test length of 60 items. This was done to more clearly demonstrate the effects of the different item parameters on scoring. Unlike reading, rescaled item parameters were not used in the simulations. The rescaled and not-rescaled parameters were very similar, which suggests that the groups were truly randomly equivalent. Two thousand examinees were simulated at each of 13 equally-spaced ability points between ±3.0. In the comparison labeled "True/True," examinees responded according to the "true" parameters (calibrated from the Pretest Order 1 administration), and were scored using the same parameters. In the comparison labeled "True/Mis," examinees responded according to the true parameters, and were scored using the mis-specified parameters (calibrated from the Pretest Order 2 administration). As with reading, only one replication was conducted. Only results for EAP estimates are presented.

Figure 10 shows the total test information by calibration group for the two pretest units (15 items). The information curves show that the set of items was much harder than the reading items. Information is at a maximum at an ability of about $\theta = 1.4$. If the testing population is normally distributed, information would be maximized for examinees that are about one to two standard deviations above average ability. Information is higher for the Pretest Order 1 group at abilities of $\theta < 1.6$. Information is higher for the Pretest Order 2 group at abilities of $\theta \geq 1.6$. The information curves show that calibrations conducted on different orderings of items can yield different results.

Figure 11 shows the average absolute bias at each true ability point for the EAP ability estimates. The True/True condition shows less bias than the True/Mis condition for abilities of $\theta \geq -1.0$. The True/True condition shows more bias than the True/Mis condition for abilities of $\theta < -1.0$. Although they are not presented here, results for the BME estimates were similar to the results for the EAP estimates (as were results for the MLE estimates excluding examinees with scores of ±5.0). Unlike the case for the reading simulation, the shift for the True/True condition from less biased than the True/Mis condition at the higher abilities to more biased than the True/Mis condition at the lower abilities, does not appear to be closely tied to the test information noted for the two sets of item parameters. It is true that the True/True case is more biased than the True/Mis case only at the low ability scores at which there is very little information available. This trend may be largely noise.

Because there is no inherent right order associated with the math pretest unit, the designation of the true and mis-specified parameters is an arbitrary assignment. The results suggest that calibrations conducted on different orderings of items can yield somewhat different results, and that if examinees take those items operationally in one order but are scored using parameters calibrated from another order, this can create unnecessary bias in their scores. If it is necessary to obtain item parameters based on pretest data, and item orders change from the pretest to operational administration, it might be best to pretest the items in more than one order. This could mitigate some of the position effects. Figure 12 shows the average absolute bias for the EAP ability estimates for the case where the true parameters were obtained from a simultaneous calibration of the data from the Pretest Order 1 and Pretest Order 2 groups. Figure 12 shows the same trends in bias as Figure 11, only the difference in bias observed across the two comparisons is reduced.

Implications for Using Pretest Item Statistics

Evaluating context effects is a difficult task, because there are many factors that can contribute to their existence. Isolating those factors in order to sufficiently study them may be impossible. This paper looked only at the effect due to differing item positions across pretest and operational administrations. How applicable the pretest-operational contexts observed here are to other testing situations is unknown. Thus, how well the results observed here generalize to other pretest units or other pretesting situations is unknown. What is relevant for one particular pretest-operational scenario is not necessarily relevant for a different pretest-operational scenario.

For the reading and math pretest units studied, there were some significant differences in item statistics when the items were administered in different positions. Whether these differences are important is a question that must be asked. Context effects can only occur when item characteristics are created from an item administered in one context, and those characteristics are then used to represent the same item when it is administered in a different context. To truly evaluate context effects, a target context needs to be clearly defined. This paper assumes that the operational administration is the desired context from which to characterize an item, and that pretesting occurs in a different context. Exactly what effect differences in context have across pretest and operational administrations likely depends on how similar the pretest context is to the context in which the items would be administered operationally, and for what purposes the pretest data are used.

In the ideal casc, if itcms arc pretested in exactly the context in which they will appear operationally, there should be no context effects. It would be an impossible task, however, to administer items operationally in exactly the same context in which they were pretested. Even if items could be pretested in the same order and with the same set of items that they would appear operationally, motivation is always likely to be a concern if appended pretesting is used as the method of pretesting. The fact that the pretest administration is not the operational administration inherently creates a different context for the two administrations.

Whether a different context across pretest and operational administrations is trivial or non-trivial is most dependent on how the pretest data are used. Item statistics computed from pretest data are commonly used to build test forms or CAT pools. Some consequences of mis-specified parameters (i.e., parameters that differ from what they would be if they were calibrated from an operational administration) in building forms/pools could be that useful items are excluded from the form/pool, forms/pools are more or less difficult than desired, or forms/pools are more or less informative than desired.

If item statistics computed from pretest data are also used for item selection (in a CAT) and/or computing IRT-based scores (for CATs or fixed-form tests), the consequences of context effects can be more costly. It is highly likely in a CAT that item parameters calibrated from pretest administrations are used in operational administrations. If there are context effects across the pretest and operational administrations, examinees could be administered inappropriate items and scored at a lower level of precision than desired. For a fixed-form test, item selection is not an issue, but examinees could still receive scores less precise than desired. Results from the simulations in this paper showed some small-to-moderate negative effects on scores when mis-specified parameters were used for IRT scoring. Larger negative effects occurred at score points where fewer examinees score. Score bias could be smaller or larger under different conditions from those observed in this study.

This study only looked at context effects for fixed-form administrations, so it is impossible to make inferences about the effect of different pretest and operational administration contexts for a CAT based on these findings. However, considering fixed-form tests versus CATs is also relevant when determining the consequences of context effects. Context effects due to item position can be clearly defined on a fixed-form test if items are forever administered in the same position. Context effects are less clearly defined in a CAT if any combination of items can precede or follow a particular item. Defining the "true" context for a particular item is difficult in a CAT. Context effects may be mitigated to some degree by administering an item in a variety of positions on the operational test, such as occurs in a CAT. Context effects may also be mitigated to some degree by pretesting an item in a variety of positions.

Another usage of pretest data is to preequate an operational test form. In this case, data from a pretest or non-operational administration is used to conduct the equating that is subsequently used for an operational administration. Kolen and Brennan (1995) distinguish between item and section preequating, and suggest that context effects need to be controlled with either type of preequating. Ideally, an operational form would be equated using data from an operational administration of that form, so that the desired administration context is obtained. This may not be a very practical practice if quick reporting of scores is desired. Results from this study support the notion that preequating should only be conducted under very carefully controlled situations. Different equating relationships could be obtained, via either classical or IRT-based methods, if there are context effects across the operational administration and the administration from which the equating is conducted.

References

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Thissen, D. (2001a). Item response theory likelihood-ratio tests for differential item functioning. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Thissen, D. (2001b). IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG. In M. du Toit (Ed.), *IRT from SSI* (pp. 24-256). Lincolnwood, IL: Scientific Software International, Inc.

Figure 1. P-Value Differences (+-2SE) for Reading Re-Pretest As Operational and Re-Pretest As Original Pretest Groups. Items With a Negative Difference Were Easier for the Re-Pretest as Original Pretest Group.

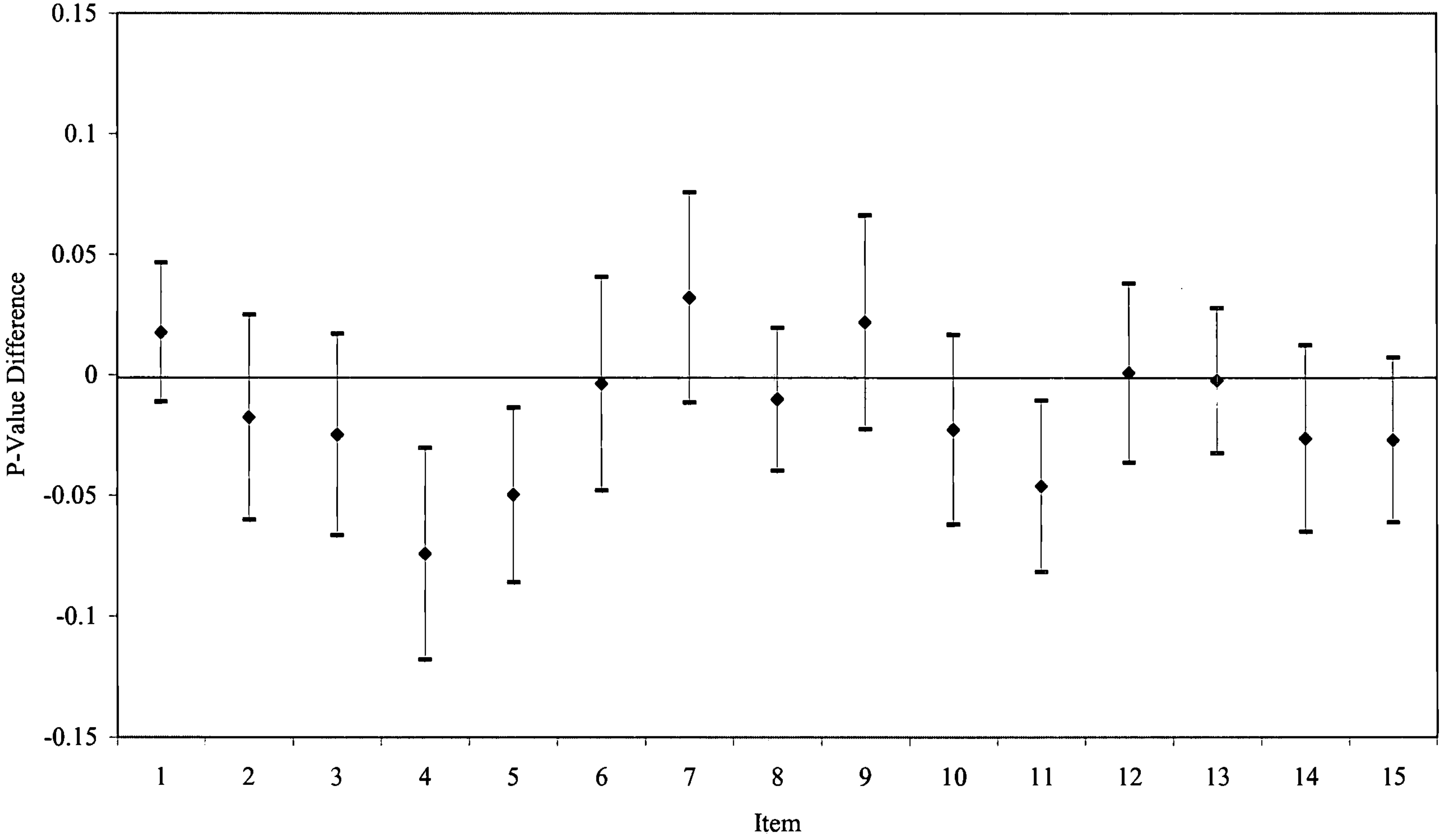


Figure 2. P-Value Difference (+-2SE) for Reading Original Pretest and Re-Pretest As Original Groups. A Positive Difference Favors the Original Pretest Group.

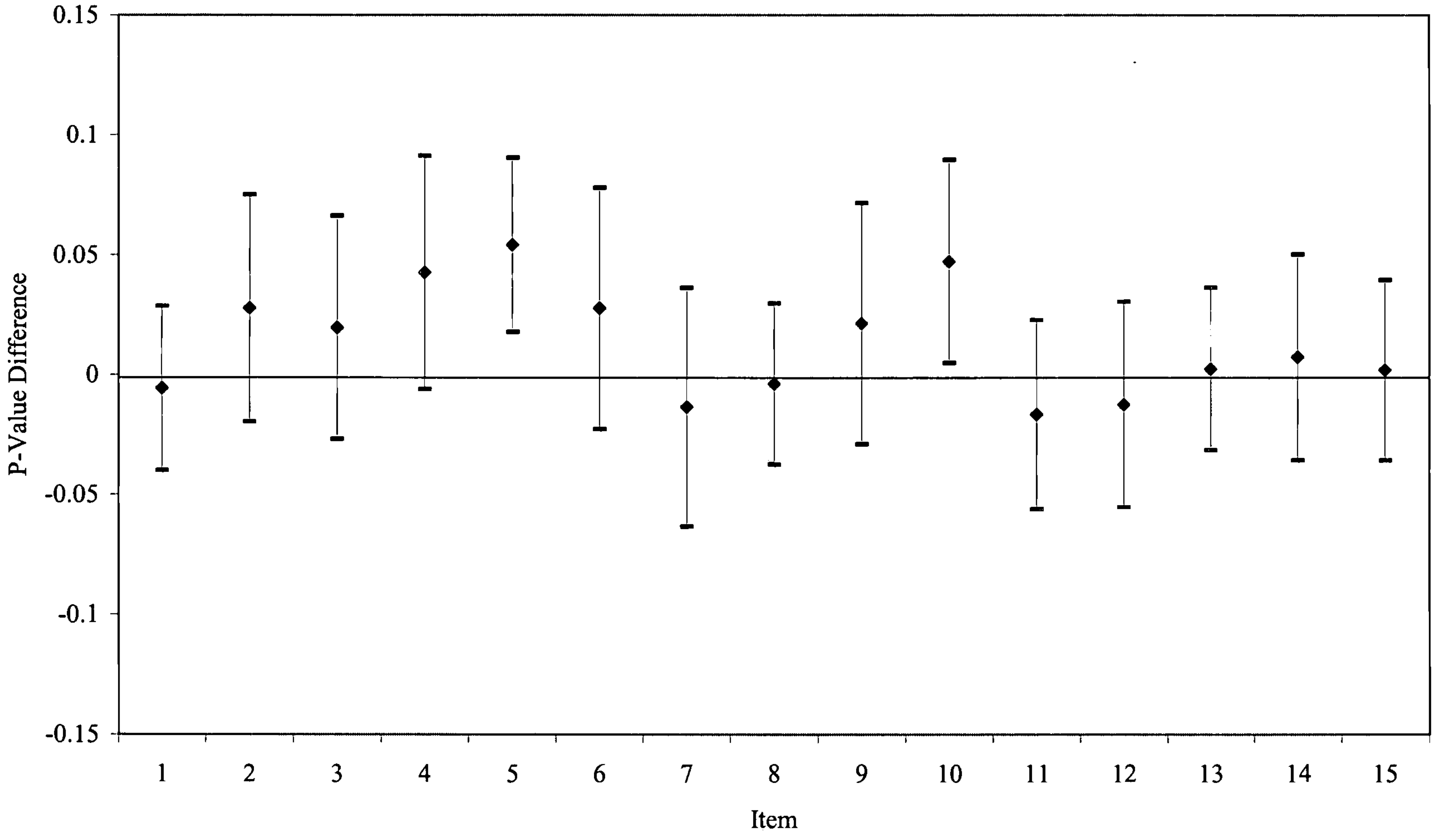


Figure 3. P-Values Differences (+-2SE) for Reading Operational and Re-Pretest As Operational Groups. A Positive Difference Favors the Operational Group.

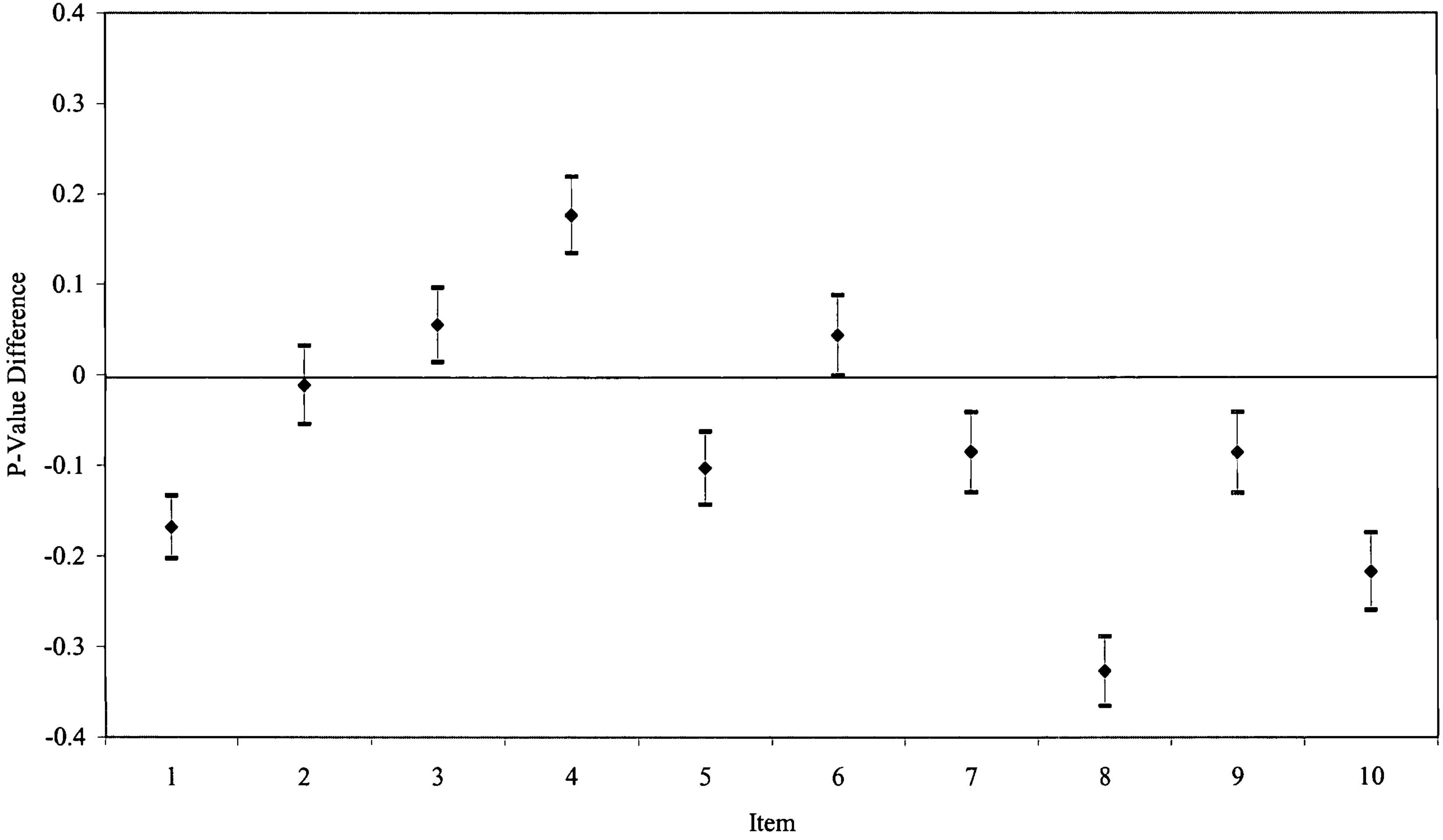


Figure 4. Total Test Information by Calibration Group for the Reading Re-Pretest Units (15 Items)

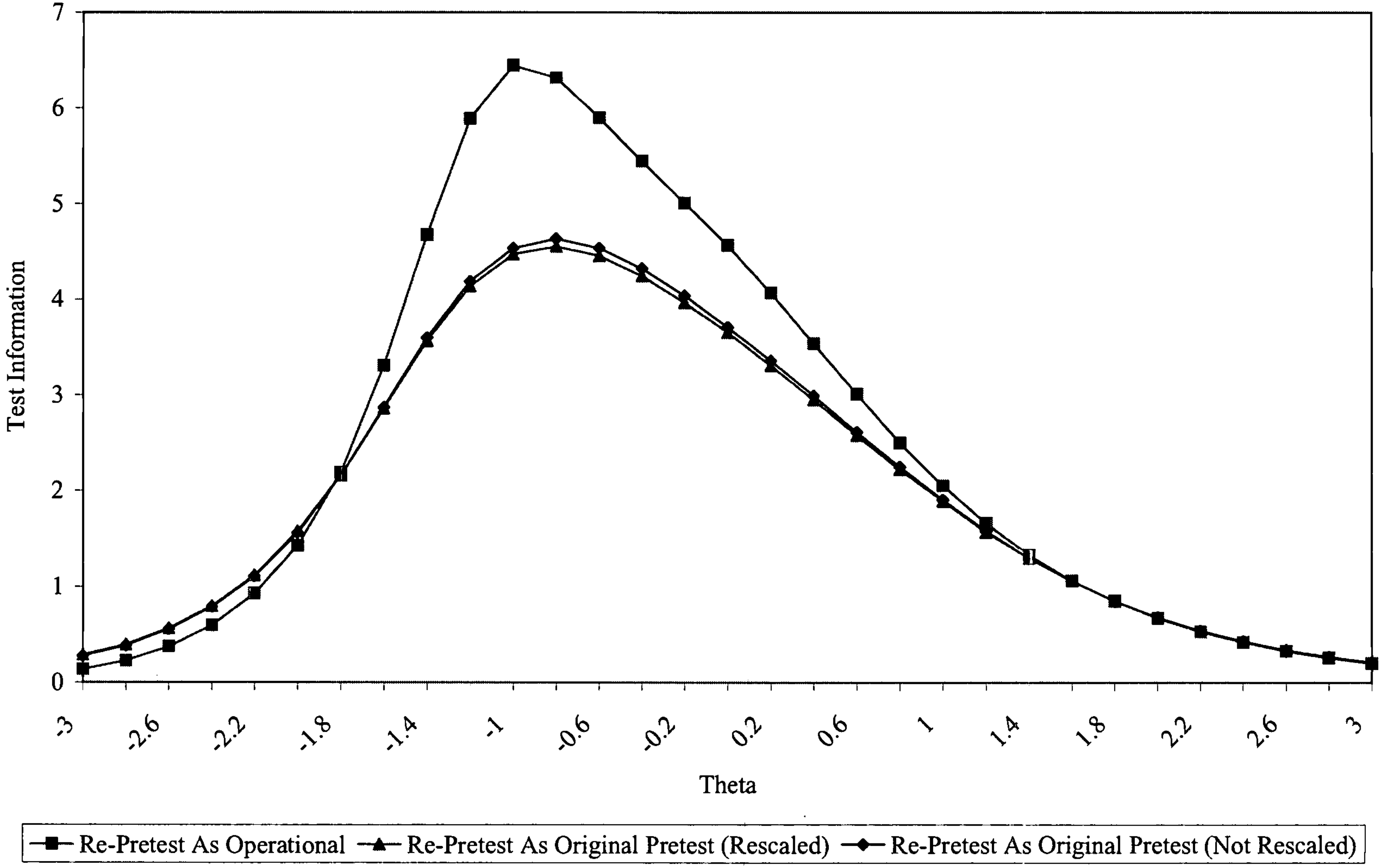


Figure 5. Average Absolute Bias for Baye's Modal Estimates, for the Reading Simulation

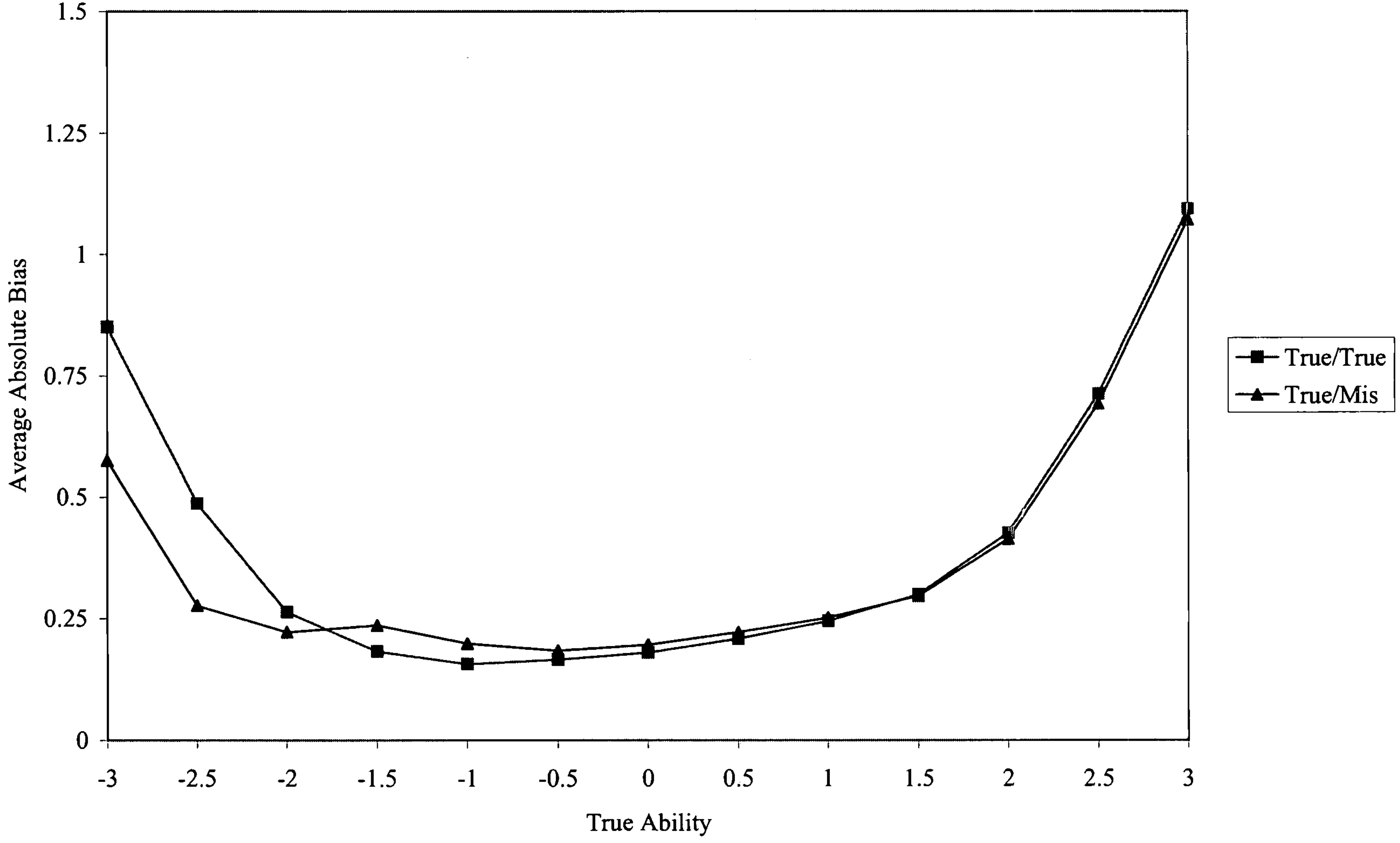


Figure 6. Average Absolute Bias for EAP Estimates, for the Reading Simulation

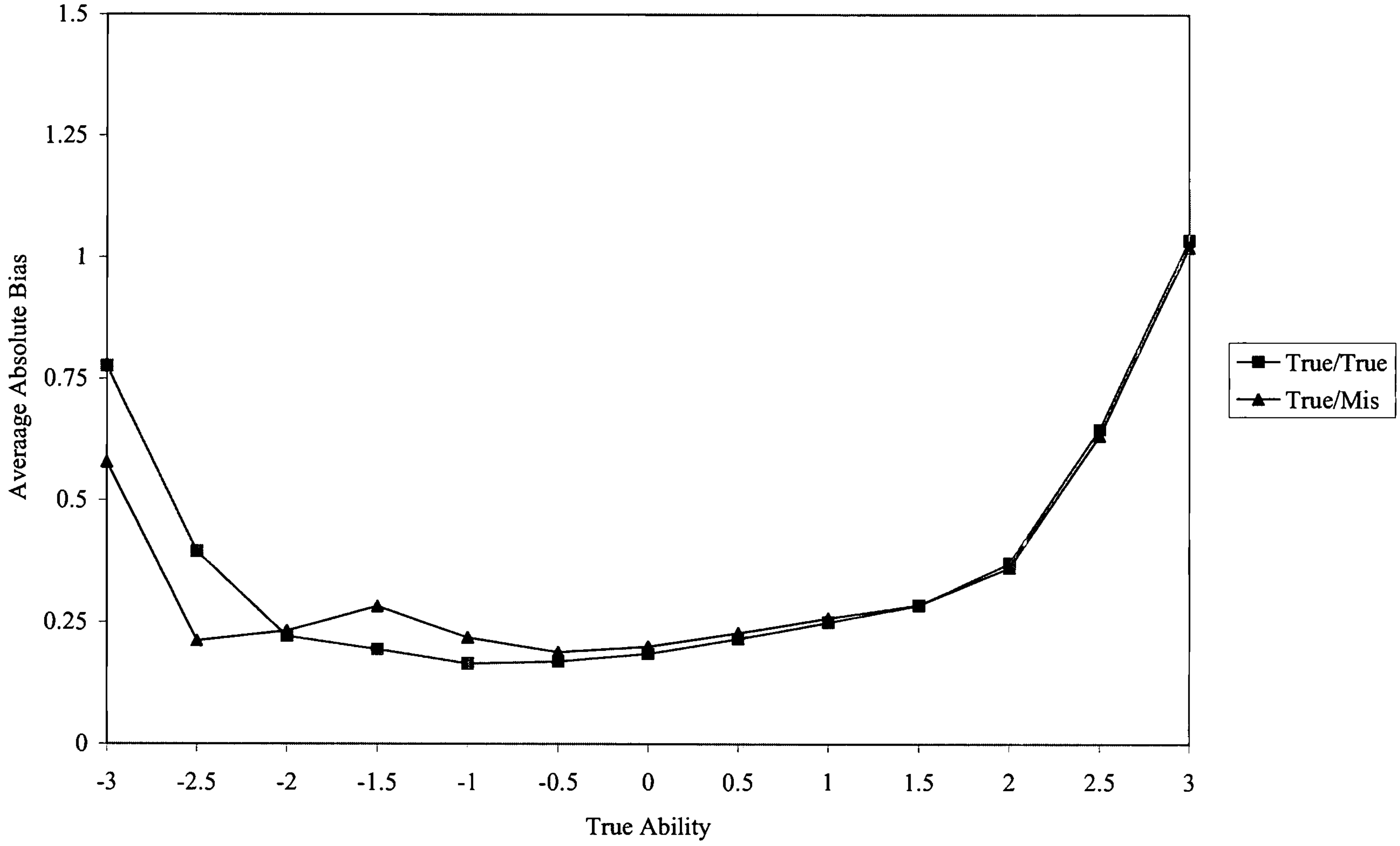


Figure 7. Average Absolute Bias for Maximum Likelihood Estimates, for the Reading Simulation

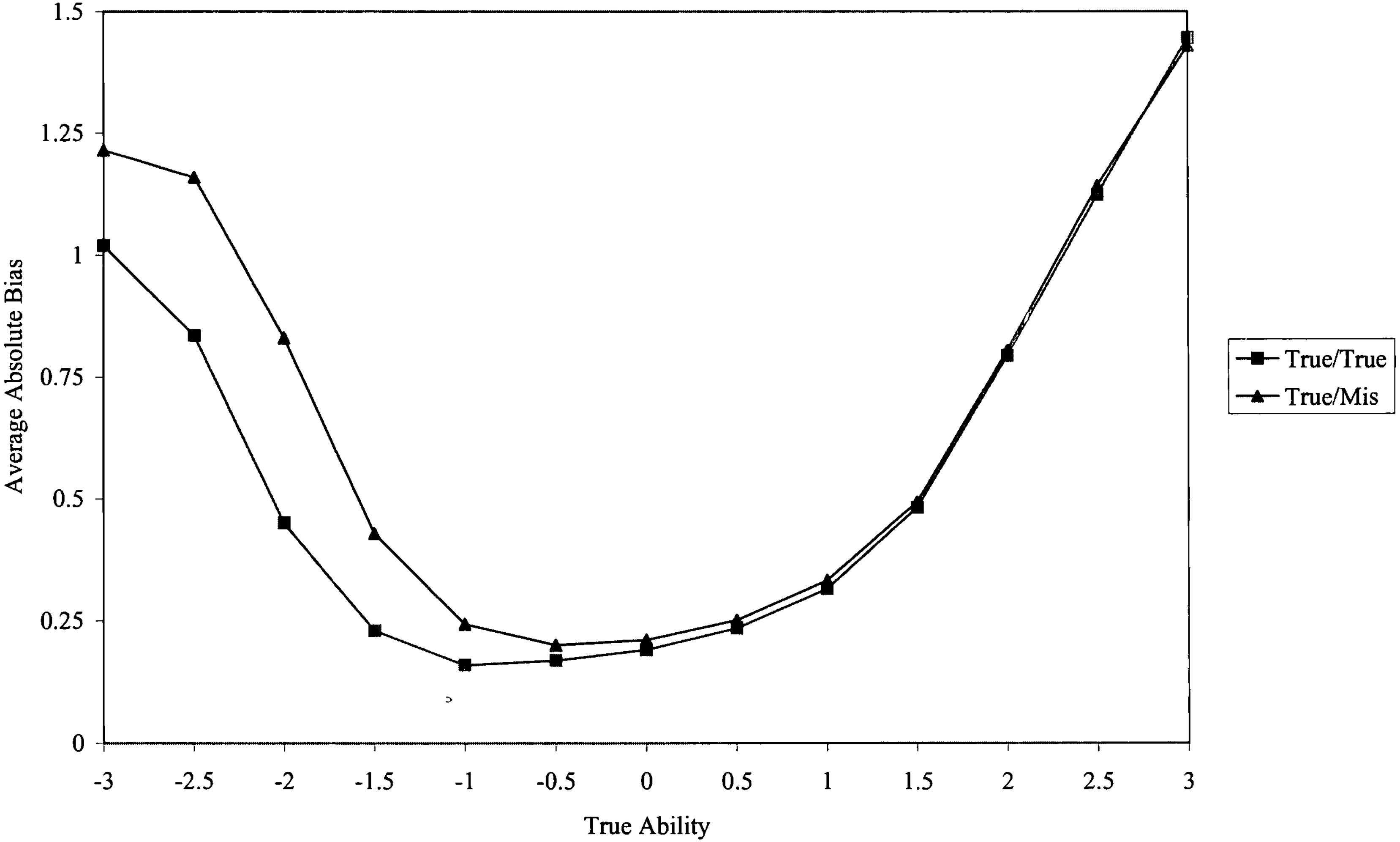


Figure 8. Average Absolute Bias for Maximum Likelihood Estimates, Excluding Examinees with Scores of +-5, for the Reading Simulation

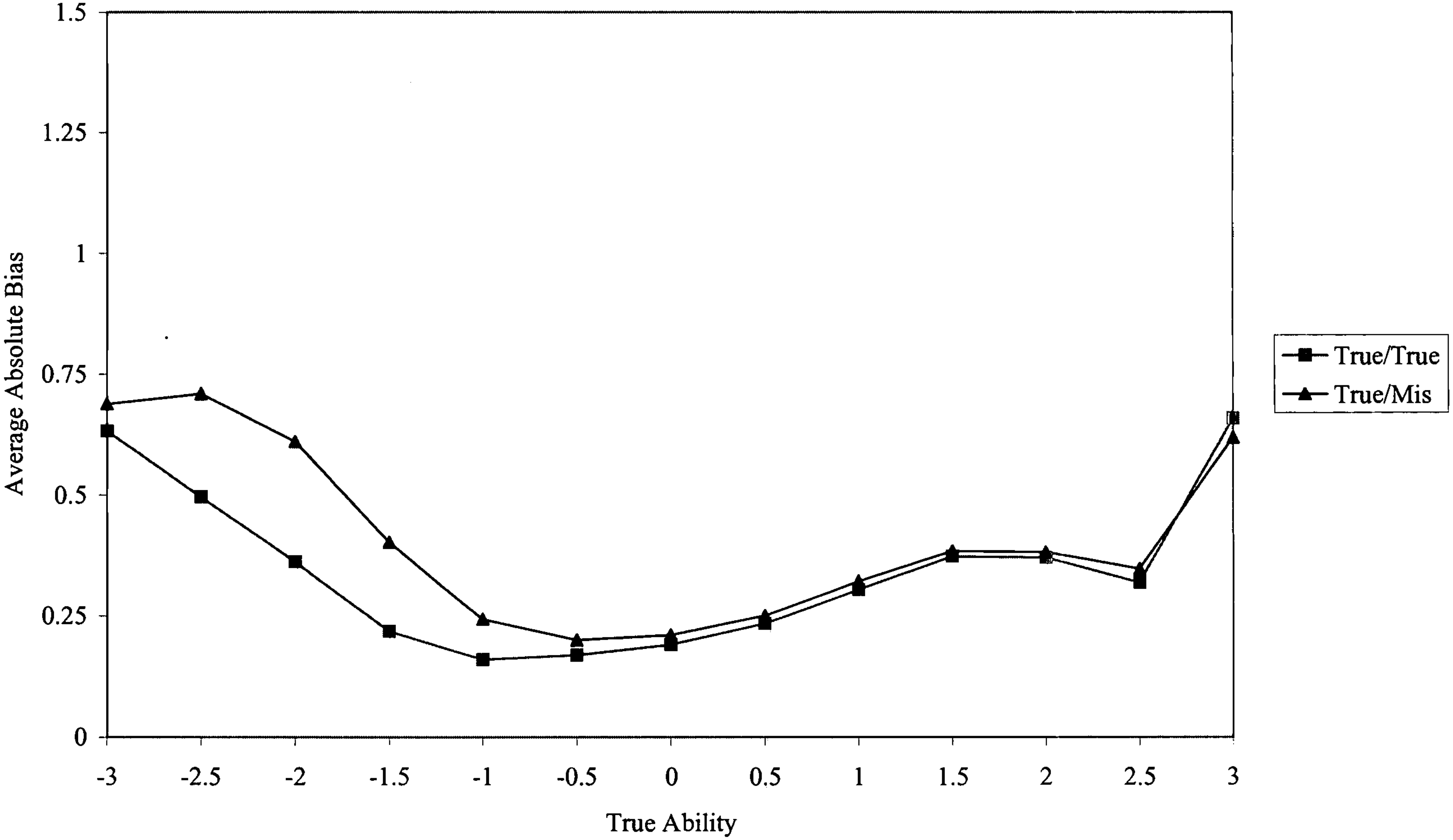


Figure 9. P-Value Differences (+- 2SE) for Math Pretest Order 1 and Pretest Order 2 Groups. Items With a Positive Difference Were Easier for the Pretest Order 1 Group.

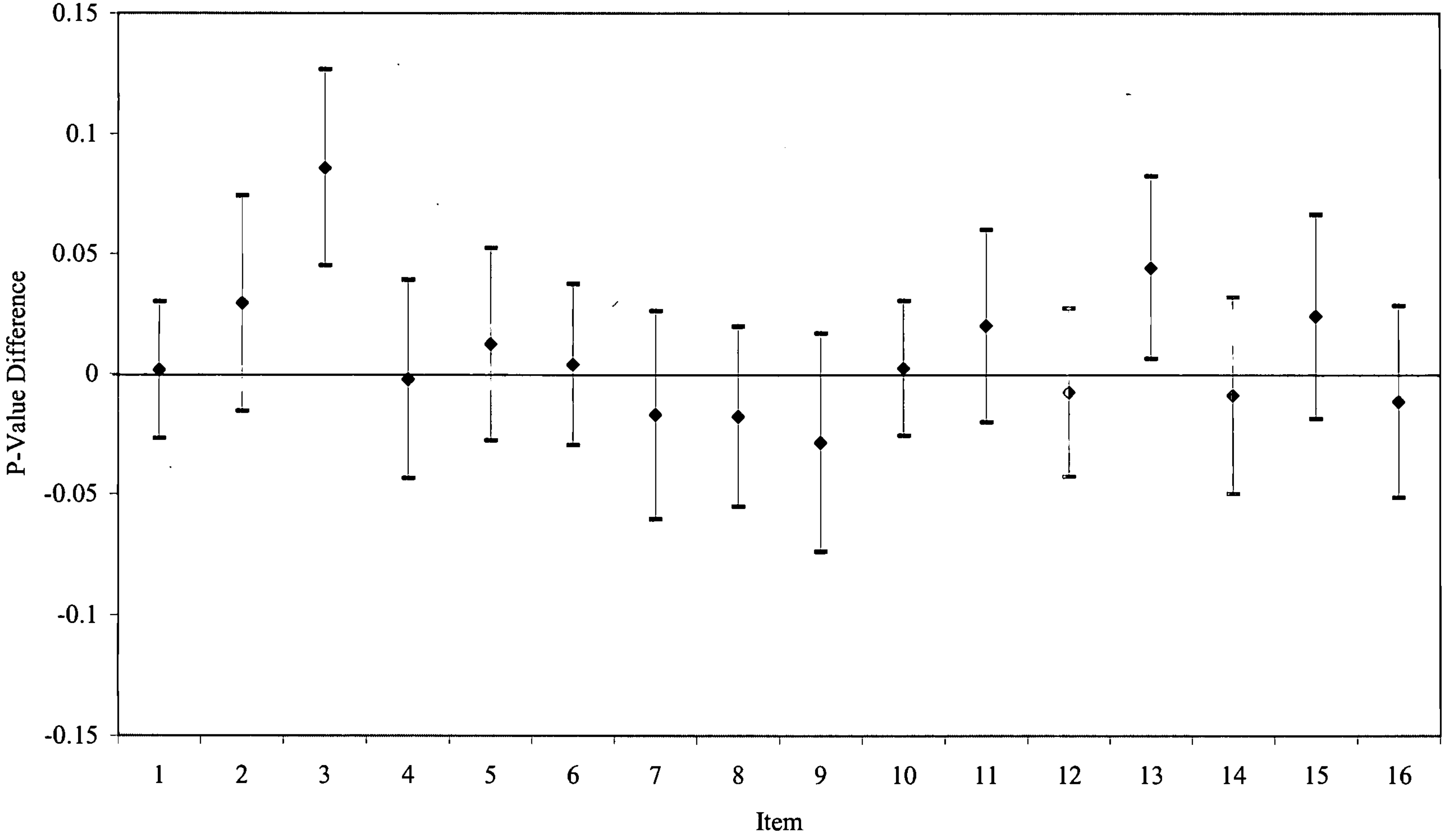


Figure 10. Total Test Information by Calibration Group for the Math Pretest Units (15 Items)

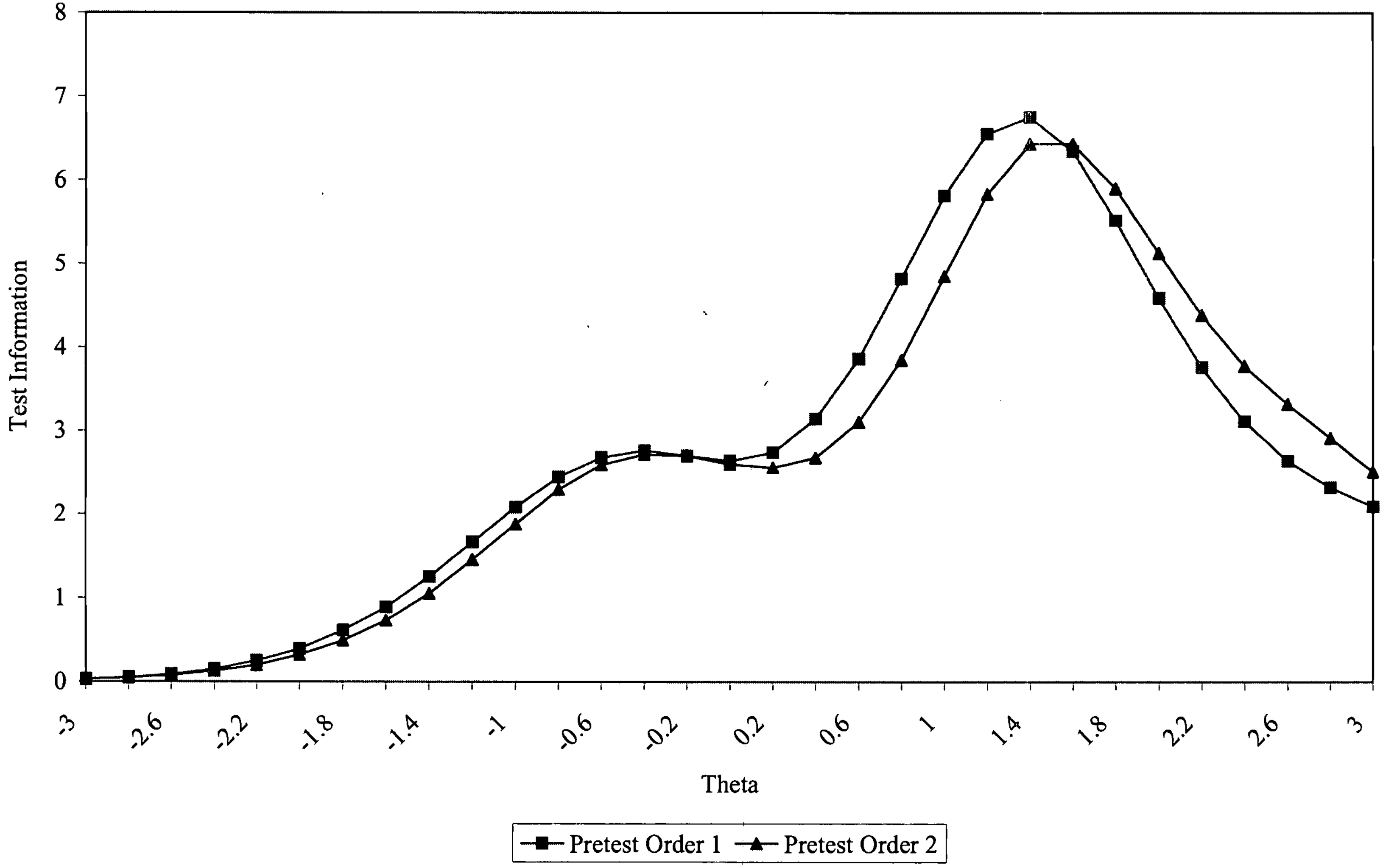


Figure 11. Average Absolute Bias for EAP Estimates, for the Math Simulation

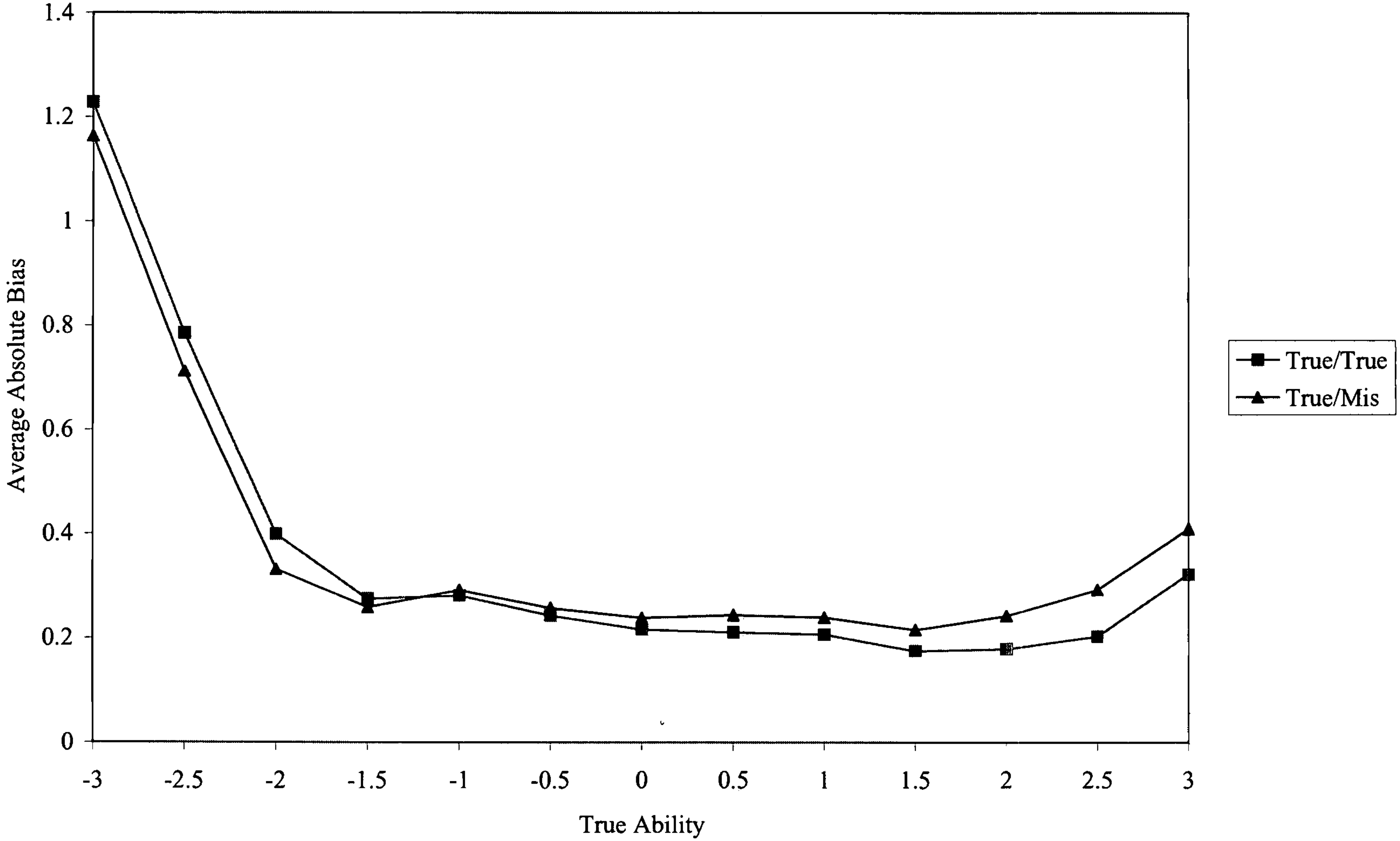
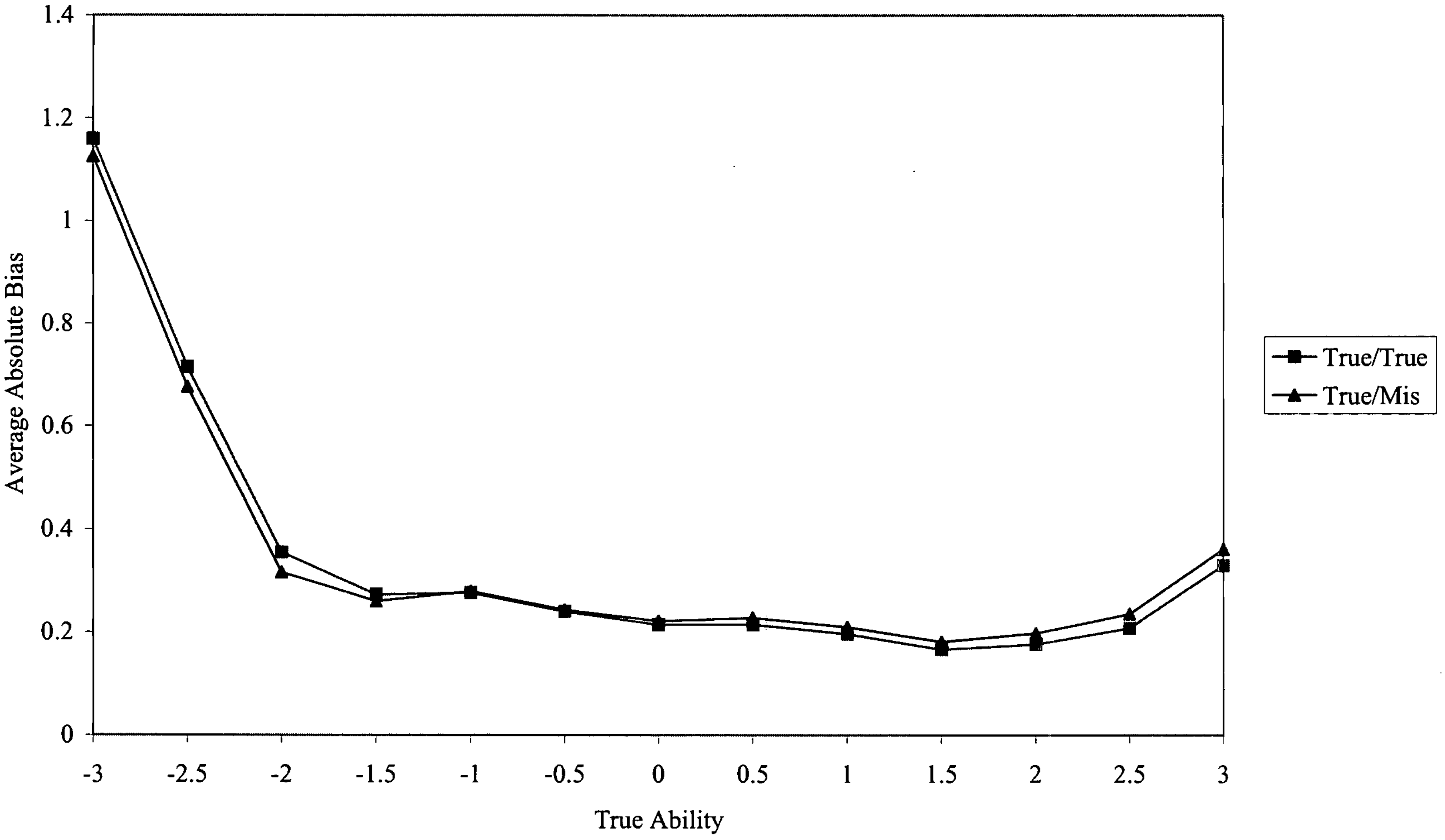


Figure 12. Average Absolute Bias for EAP Estimates Where True Parameters are From a Simultaneous Calibration of Pretest Order 1 and Pretest Order 2 Data.



REPRODUCTION RELEASE

(Specific Document)

TM034974

I. DOCUMENT IDENTIFICATION:

Title: Context Effects in Pretesting: Impact on Item Statistics and Examinee Scores	
Author(s): Mary Pommerich, Deborah J. Harris	
Corporate Source: Defense Manpower Data Center ACT, Inc.	Publication Date: April, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Mary Pommerich	Printed Name/Position/Title: Mary Pommerich / Psychometrician	
Organization/Address: Defense Manpower Data Center 400 Gigling Rd. Seaside, CA 93955-6771	Telephone: 831-583-4066	FAX: 831-583-2340
	E-Mail Address: pommie@att.net	Date: 5-22-03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org