Achieving Accuracy of Pretest Calibration for

A National CAT Placement Examination with a Restricted Test Length[1]

Xiang Bo Wang
Andrew Wiley

Research and Psychometrics
The College Board

Paper to be presented at 2004 NCME Conference, San Diego, California.

[1] The authors wish to thank Vantage Learning for implementing item pre-testing and collecting the pretest data, and Ashley Buonasera for his summer internship work on part of the project.

Introduction

For the past fifteen years, the majority of conference presentations and journal publications on computer adaptive testing (CAT) have been based on large-scale aptitude and achievement examinations which tend to be long (> 50 items) in length and administered to thousands of examinees. The Graduate Record Examination and Test of English as a Foreign Language are examples of these types of examinations. To date, there have been relatively few psychometric studies about CAT as it is applied to placement examinations. Except for the College Board's Advanced Placement Examinations, most course-related placement examinations are often short (< 30 items) and taken by only a small percentage of the general student population in any school.

The main objective of a placement examination is to provide diagnostic information about the relative strengths and weaknesses of an examinee in a particular subject area. Mostly used by school counselors, admission officers or course professors, placement testing has several unique features. First, it has more varieties: in any school or college, virtually every foundation subject, such as math or English, has a placement test. Second, a placement test must be convenient to administer and score, because school counselors, admission officers and professors often need to test students immediately or on a short notice, and prefer to receive immediate feedback to make placement decisions. Third, because of the tight schedule of school counselors and the need for admission officers to meet with many students in a short time, especially at the beginning of a new semester, a placement test must be short (often less than 15 minutes) but reasonably accurate. Fourth, placement testing is often considered a low test security risk, because most new students do not know each other intimately enough to share placement test items when they first enter a new school or class. Often, students do not even know that they will be required to take a placement examination[2]. In addition, there is little incentive for them to cheat on a placement examination as compared to other higher stakes examinations. Very few students want to be placed into a class markedly above their current skills and abilities that will lead to a high probability of class failure. These four features make CAT a highly viable delivery system for placement exams, because CAT offers the logistic flexibility and convenience as well as the measurement efficiency and accuracy as required above (Wainer,1990; Embretson, 2000).

However, the “double-edged” requirement of a short but accurate placement test poses some special challenges to CAT implementation. First, when the test length is short, measurement precision can be significantly reduced. In addition, the short test length requirement severely limits the number of items that can be seeded for pre-testing. The purpose of this study is to summarize how to best calibrate pretest items for a national CAT placement examination with a restricted test length.

The ACUPLACER™ placement examination system was jointly developed by The College Board and Educational Testing Service. The ACUPLACER™ system has a number of individual examinations within the overall system. One such exam, the ACCUPLACER™ Elementary Algebra exam (AEA), is a placement examination used

[2] International students whose native language is not English may be the only exception, since they are often notified of the need to take an English placement examination when they first enter an American university or college.

by community and four-year colleges throughout the nation to assess the adequacy of a student's knowledge in elementary algebra for the purposes of class assignments, assessment of academic progress and/or evaluation for remedial classes (The College Board, 2000a, 2003). In order to satisfy the needs of school counselors and admission officers, AEA is not only computer adaptive, but also delivered over the Internet, so administration can be on demand through secured authorization anytime and anywhere. The Internet based delivery system has also made it possible to centrally monitor test quality and security, and to update item pools and other components of the examination system. Studies have shown substantial predictive and concurrent validities for the ACCUPLACER™ assessment system including AEA (The College Board, 1999a, 1999b, 2000C).

As stipulated by test specifications, AEA is an un-timed fixed-length CAT with 12 operational items. Based on the three parameter logistical (3PL) IRT model (Birnbaum 1968), AEA uses the weighted deviation model (Stocking & Swanson, 1993) for item selection. Between June 2000 and April 2001, a batch of over 473 new items was pre-tested to replenish the item pool. The goal of this study is two-fold; to summarize the results for pre-test calibration and scaling under the restrictions described above; and to describe the results of one of the three simulation studies designed to evaluate the accuracy of the observed pre-test calibration results.

Research Design

Part 1: Real-Data Calibration

Data Collection:

The data used to calibrate the pretest items came from 364,018 examinees who took the AEA test between June 2000 and April 2001. After data cleaning, the final version of this data set contained 151 operational items and 473 pretest items in a 624 by 364,018 square, but spare matrix. The sparseness of the data matrix was a result of the short length of the test, with only 12 operational and 5 pretest items. Each examinee was presented with a different set of 17 items. The remaining 607 items that an examinee did not take were represented as not presented.

Pre-testing Plan:

After careful discussion, the AEA test committee decided that the number of pre-test items given to any test taker would be five in order to maintain a balance between the brevity of the test and the need to pretest all 473[3] pre-test items with a minimum of 1,000 examinees. To avoid significantly affecting the flow and accuracy of an already-short test, the AEA test committee also decided to embed the five pre-test items within the CAT session. Examinees were not explicitly informed which five items on the test were

[3] Originally, a total of 475 items were pretested. Two pretest items were excluded from this report due to item accession number mismatch.

pre-test items and would not contribute to his/her score until after they received their results. Furthermore, to avoid the potential impact of restricted ability range on item calibration as often seen in on-line calibration data (Ito & Sykes, 1994; Sykes & Ito, 1995), all the pretest items were randomly administered to examinees of all ability levels instead of restricted ability ranges. The pre-testing data collection went on for nearly one year and a giant sparse response matrix of over 364,018 examinees was obtained.

Calibration Plan:

In order to minimize errors introduced by multiple stages of calibration (Wightman & De Champlain, 1994), a concurrent calibration scheme was used which included all the 151 operational and 473 pre-test items along with the sparse response matrix of 364,018 examinees. The BILOG-MG Version 3.0 (Scientific Software Inc. 2003) was used to carry out the concurrent calibration, since research has shown that BILOG was able to recover item parameters better than LOGIST in terms of shorter tests and smaller examinee sizes (Yen, 1987; Tang and et al, 1993) and sparse matrices (Smith, Rizavi, Paez, & Rotou, 2002; Ban, Hanson, Yi, & Harris, 2002).

Scaling Plan:

The mean/sigma method (Marco, 1977) was used to put the above calibration results on the same scale as those of the original item pool. Note that the original parameters of the 151 operational items were obtained from paper and pencil (P&P) administrations using the LOGIST program (Wingersky, 1983). Due to a variety of reasons, neither the original item response data nor the LOGIST calibration program was available to the authors at the time of this study.

Results from Real Data Calibration

The results of the BILOG item calibration can be summarized across five broad areas. The five areas that will be summarized are: 1) Findings on item fit in BILOG CAT calibration, 2) Findings on item parameter estimates between LOGIST and BILOG, 3) Findings on item characteristic curves between LOGIST and BILOG, 4) Findings on test characteristic curves between LOGIST and BILOG, and 5) Finding on test information curves between LOGIST and BILOG.

Findings on Item Fit in BILOG CAT Calibration:

Given the extreme sparseness of the item response data matrix and the short test length of each CAT session, the first question to be answered is the extent of item fit from the concurrent BILOG calibration. The first finding is that 470 out of the total of 473 pretest items were fit very well by the 3PL model. This was most likely due to the fact that they were administered randomly to examinees of the entire ability range. Item fit was determined by examining item characteristic curves (ICC), observed percentage correct, and 95% confidence bands around ICCs as provided by the BILOG program. Figure 1 demonstrates the ICC and the item fit boundaries of one typical pretest item from the BILOG output.

As for the 151 operational items, most of them demonstrated adequate item fit according to the graphical representation of the BILOG program. Figure 2 shows a well-fit operational item. However, about 20 (out of 151) operational items were not fit very well. Figure 3 illustrates one poorly fit operational item. Despite the fact that over 11,000 examinees responded to this item, it is clear that the main cause of the item misfit was the extremely restricted ability range of the candidates who saw each item. This finding confirmed the phenomenon reported by Ito & Sykes (1994). After a thorough evaluation by psychometric and content staff, it was decided that ten items that had severe item misfit would be removed from the pool of anchor items being used for scaling the pretest items.

Findings on Item Parameter Estimates between LOGIST and BILOG:

After removing the 10 severely misfit items, the remaining 141 operational items were scaled to their LOGIST counterparts using the mean-sigma method. Table 1 summarizes the descriptive statistics on the parameter estimates of BILOG and LOGIST. While the BLOG and LOGIST *b* parameter estimates are expected to have identical mean and standard deviations using the mean-sigma scaling method, both the *a* and *c* parameter estimates also had highly similar mean and standard deviations.

Table 1: Summary of Item Parameter Estimate between BILOG and LOGIST

Parameters	N	Mean	Std Dev	Minimum	Maximum
BILOG A	141	1.395	0.623	0.308	3.688
BILOG B	141	0.241	0.809	-1.994	1.714
BILOG C	141	0.153	0.085	0.007	0.427
LOGIST A	141	1.344	0.396	0.383	2.114
LOGIST B	141	0.241	0.809	-1.698	2.431
LOGIST C	141	0.184	0.087	0.015	0.500

The correlations between the BILOG and LOGIST *a*, *b* and *c* parameter estimates of the 141 operational items were substantial. The correlation was 0.94 between the *b* parameters and 0.71 for *a* parameters. The correlation between the BILOG and LOGIST *c* parameters was moderate, only 0.56. Such a correlation pattern was typical of CAT data (Tang and Eignor, 1997).

Figures 4 to 6 show the scatter plots of the *a, b* and *c* parameter estimates between LOGIST and BILOG after the mean-sigma scaling, the regression line and 95% confidence intervals. Three patterns are clear. First, as shown by Figure 4, the BILOG CAT calibration seemed to have produced higher *a* parameter estimates than the LOGIST P&P calibration. Second, the *b* parameter estimates for the great majority of items were highly similar. Third, more items had slightly lower *c* values from BILOG CAT calibration than from LOGIST P&P calibration. These three patterns correspond to the findings from some previous research (Tang & Eignor, 1997).

Findings on Item Characteristics Curves between LOGIST and BILOG:

The correlations between LOGIST P&P and BILOG CAT parameter estimates were also reflected in the similarities and differences of the ICCs of the two calibrations. Figure 7 shows an item whose BILOG and LOGIST ICCs were virtually identical, while Figure 8, illustrates an item whose BILOG and LOGIST ICCs were most different among the 141 operational items. The ICCs of the other items varied between these two exemplary ICCs. Based on visual evaluation, the ICCs of 35 items resembled those in Figure 7, while the ICCs of 16 items were similar to those in Figure 8. The ICCs of the remaining items looked acceptably close.

Findings on Test Characteristics Curves between LOGIST and BILOG:

The next area investigated how similar the test characteristic curves (TCC) between LOGIST P&P and BILOG CAT calibrations were. On the basis of the entire 141 operational items, Figure 9 shows that the two TCCs were highly similar and parallel. Across most of the ability range, the LOGIST TCC was slightly higher than that of BILOG TCC. As indicated by Figure 10, the biggest difference was approximately 4 points, which occurred around ability points –4.0 and –0.5. The significance of the high TCC similarity lies in the fact that if IRT true score equating is employed, the biggest expected score difference between taking the 141 items from BILOG CAT and LOGIST P&P calibrations, respectively, would be only four points. The difference would be much smaller under the current AEA CAT test lengths of 12 operational items!

Findings on Test Information Curves between LOGIST and BILOG:

The last area investigated how similar the test information (TIF) curves between LOGIST P&P and BILOG CAT calibrations were. According to Figures 11 and 12, the BILOG CAT calibration yielded higher test information than the LOGIST P&P parameter estimates across most of the middle ability range, namely between –1.0 and +1.0. The biggest TIF difference was about 14 points. This higher information could be related to the fact that in CAT test sessions, items were selected based on the optimal test information.

Summary on Real Data Calibration and Next Steps

Based on the findings summarized so far, two major conclusions can be made. First, the item parameters of the 473 pre-test items seemed to have been estimated as appropriately and accurately as they could have been due to two reasons. First, all the pretest items were administered randomly to sufficient numbers of examinees throughout the entire ability range. As a result, there is little interpolation by BILOG in fitting the ICC curves for the pretest items, because all of them had adequate response information for calibration across the entire ability[4]. In essence, calibrating these pretest items is no different from calibrating pretest items from regular P&P administrations with a lot of non-presented items!

[4] The general rule of thumb is to use about 800 examinees per item as a sound examinee sample size.

Second, the item parameter estimates of the 141 operational items were adequate from the standpoints of item fit and substantial similarity with their original LOGIST parameter estimates, and especially test characteristic curves. The relative small TCC differences between the LOGIST P&P and the BILOG CAT calibrations were specially encouraging because they meant practically little or no differences in average equated scores between LOGIST and BILOG calibrations.

Based on the authors' past experiences from working on large-scale CAT examinations and CAT calibration literature, some differences were expected in the ICCs between LOGIST P&P and BILOG CAT calibrations. Such differences can be attributed to the combination of three factors. The first factor is the testing mode difference. The original LOGIST parameter estimates were obtained from P&P data in which all the items were answered by all examinees. The current BILOG data came from CAT sessions where various items were administered to different and sometimes very limited ability ranges. During the BILOG CAT calibration, a substantial amount of interpolation was employed to fit the ICCs for the ability ranges that had very little examinee response information. The second factor causing the parameter estimate and ICC differences could be the algorithmic differences between LOGIST and BILOG programs (Mislevy & Stocking, 1989). The former uses the conditional maximum likelihood estimation, while the latter, the marginal maximum likelihood estimation. The third factor could be the differences in test lengths. The original P&P test length for LOGIST calibration was 120 operational items, while the current CAT test lengths for BILOG calibration were only 12 operational items.

The question to be further investigated by this study is to what extent each of these three factors could have influenced the observed differences between LOGIST and BILOG calibrations. Unfortunately, the original LOGIST calibration data is no longer available. If it were available, we could have assessed the effects of the second factor, the effect of the algorithmic differences between LOGIST and BILOG. Due to the copyright and licensing restrictions associated with the LOGIST program, we could not have used it for this project either. The only choice left is to carry out simulation studies to assess the effects of the three factors on the observed differences between BILOG and LOGIST parameter estimates.

Part 2: Summary of Simulation Designs and Results of Simulation 1

To sort out the effects of the factors mentioned above, three types of simulations seemed appropriate. In this part, we will summarize the design and results of the first simulation. In the section on future research, we will describe the designs for Simulations II and III.

Design on Simulation I: BILOG Calibration under Ideal Linear Data Condition.

According to Mislevy and Stocking (1989), an ideal data condition for a BILOG calibration is a full data matrix which contains response data to a minimum of 50 items, and each of these items is responded to by, at least, 1,000 examinees whose abilities are normally distributed. This scenario exemplifies a typical linear P&P data condition. It is well known in the IRT calibration literature that the item parameter estimates obtained from such an ideal calibration condition would produce the most reliable and accurate item parameter estimates. The fundamental purpose of this simulation was to establish a benchmark of variation of ICC to assess the extent to which the item parameter estimates and the ICCs from the observed real data BILOG CAT calibration differed from their expected values.

To satisfy the ideal data conditions, the LOGIST item parameter estimates for the 141 operational items were treated as if they had been true item parameters, and 5,000 examinees with normally distributed ability were simulated. The item responses of the 5,000 simulees to the 141 items were generated using the common IRT response simulation method. This method compared the probability in which a simulee would answer each item correctly with a random uniform probability. If the former was larger than the latter, a correct response of “1” was assigned. Otherwise, an incorrect response of “0” was generated.

The above simulation process was repeated 100 times. Each round of simulation would generate a different set of normally distributed simulees and a new 141-item-by-5,000 simulee response matrix. Note that throughout the 100 rounds of simulations, the item parameters remained constant. In order to obtain item parameter estimates, 100 rounds of BILOG calibrations were performed on the 100 sets of simulated response data.

At the end of the 100 rounds of BILOG calibrations, 95% confidence intervals were established for the ICCs of the 141 items. If the observed ICCs from the real data BILOG CAT calibration fell within the 95% confidence intervals, it can be concluded that the observed BILOG CAT parameter estimates and their ICCs were not significantly different from their original LOGIST counterparts. It can be further concluded the three factors discussed previously would not have any significant impact on the BILOG CAT calibration. However, if the findings showed that the observed BILOG CAT ICCs fell outside the 95% confidence intervals, it can be concluded that the observed BILOG CAT parameter estimates are significantly different from their original LOGIST P&P parameter estimates under the ideal data conditions. Note that the negative findings from this simulation would not explain the potential influences caused by short test lengths or CAT data, both of which were to be answered by Simulations II and III.

Summary of Results from Simulation I

In general, the simulation under the ideal linear data condition produced very narrow 95% confidence bands for the ICCs across 100 rounds of simulation, and consequently, the close target and average ICCs were extremely close. Figure 13 shows one of the typical tightest ICC confidence bands, while Figure 14, the widest ICC confidence band among all the 141 operational items. A visual inspection of the 141 operational items would reveal that fewer than 10 items had the spread-out ICC confidence bands close to that in Figure 14.

The next area investigated was the extent to which the ICCs of BILOG item parameter estimates calibrated under the real AEA CAT data conditions fell within or approximated the ICC confidence interval bands from the simulations. Figure 15 shows an example of an exact “fall-within”; Figure 16, a close approximation; and Figure 17, a substantial deviation. Again, using visual inspection, there were 14 items whose ICCs fell into the simulated 95% ICC confidence intervals as in Figure 15, while the ICCs of 11 items deviated substantially, but less than that of Figure 17. The ICCs of the remaining 116 items approximated their corresponding simulated ICC confidence bands as in Figure 16.

Concluding Remarks

Based on the results from Part I on real data calibration and scaling, two main conclusions can be made. First, the pretest items were adequately seeded in the AEA CAT administrations, and their parameter estimates seemed to have been satisfactorily calibrated. Second, the majority of the 141 operational items were also well calibrated, although varying ICC differences were observed for a small number of items between their LOGIST P&P and BILOG CAT calibrations. Although such differences were expected and often accepted as legitimate and common for practical situations, this study went further to investigate what could have caused such differences. Results from Simulation I demonstrated that these differences, for the most part, were larger than random calibration variations under ideal data conditions. Two additional simulations are under way to verify if these observed differences would be larger than those introduced by shorter test lengths and/or CAT testing conditions.

The significance of this study lies in the fact that it is one of the few studies that attempts to fully account for, through experimentally designed simulations, the ICC differences possibly caused by a combination of three factors: LOGIST vs. BILOG algorithmic differences, long vs. short test lengths, and P&P vs. CAT testing modes. It is hoped that by the time Simulations II and III (to be described in the next section) are completed, this research will prove to be beneficial within a variety of contexts. First, it will establish the legitimacy of the methods employed in this study to obtain accurate pre-test item and examinee parameter estimation in the context of a CAT placement examination when the test length is very limited. Both practical experiences and theoretical inferences can be gained from both the real-data application and simulations. Second, the results from this study also provide information regarding the stability and accuracy of BILOG calibration on small numbers of items in the CAT context. Third, the study also provides useful information regarding the BILOG to LOGIST scale conversion under the restricted CAT situation.

Future Research

In order to fully account for the differences in the parameter estimates between their BILOG CAT and LOGIST P&P calibrations of the 141 AEA operational items, two additional simulations will be pursued. Simulation II is planned to investigate whether or not the remaining ICC differences from Simulation I can be explained by the possibility that they were caused by much shorter test lengths. As shown earlier, the number of items in Simulation I were uniformly 141 items, while the operational number of items of

AEA CAT was only 12 items. It is reasonable to speculate that such a substantially shorter test length would impact on the stability of item parameter estimates, consequently larger ICC confidence bands. The outline for Simulation II is to randomly administer only 12 items (out of the 141 operational items) to simulees. The same simulees and item parameters from Simulation I will be used. One hundred rounds of simulations and BILOG calibrations are to be carried out, and 95% ICC confidence interval bands will be constructed to evaluate if they would embrace the ICCs from the operational BILOG CAT calibration.

If the results from Simulation II are still not 100% positive, Simulation III will be conducted to ascertain if the observed BILOG CAT item parameter estimates and/or ICCs would fall within the 95% confidence intervals under multiple CAT data calibrations. Again, the LOGIST P&P parameter estimates would be treated as true parameters, and the same 100 groups of 5,000 simulated examinee abilities used from Simulation I would be treated as true abilities to simulate CAT responses. All the realistic operational AEA CAT administration rules would be applied, including content constraints, item exposure and test lengths. Again, 100 rounds of BILOG calibrations would be run on the simulated responses, and the 95% confidence intervals of item ICCs would be established.

It is hoped that the completion of these two additional simulations will allow for a more complete explanation of the differences between the LOGIST P&P and the BILOG CAT parameter estimates.

References:

Birnbaum, A. (1958). Some latent trait model and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Ban, Jae-Chun, Hanson, Bradley, Yi, Qing and Harris, Deborah, (2002): Data sparseness and online pretest item calibration/Scaling Methods in CAT. ACT Research Report Series.

Davey, T., Pommerich, M. & Thomson, T. (1999). Pretesting alongside an operational CAT. Paper presented at the annual conference of the National Council on Measurement in Education in Montreal, Quebec, Canada.

Ito, K. & Sykes, R. (1994). The effect of restricting ability distributions in the estimation of item difficulties: implications for a CAT implementation. Paper presented at the annual conference of the National Council on Measurement in Education in New Orleans.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.

Mislevy, R. J. & Stocking, M. L. (1989) A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement.* Vol. 13(1) 57-75

Smith, R., Rizavi, S., Paez, R. & Rotou, O. (2002). Updated item parameter estimates using sparse CAT data. Paper presented at the annual conference of the National Council on Measurement in Education in New Orleans.

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

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

Sykes, R. and Ito, K., (1995). Estimation of item difficulty from restricted CAT calibration samples. Paper presented at the annual conference of the National Council on Measurement in Education in San Francisco.

Tang, L & Others. (1993). The effect of small calibration sample sizes on TOEFL IRT-based equating. ETS Technical Report.

Tang, L. & Eignor, D. (1997). An investigation of the relationships between IRT parameter estimates using LOGIST and BILOG. A paper presented at the annual meeting of NCME in Chicago.

Wightman, L. & De Champlain, A. (1994). A comparison of the properties of IRT parameter estimates using two different calibration designs. ETS Research Report.

Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*. Vol 52(2) 275-291.

The College Board, (1999a). Comparison of ACCUPLACER and the academic assessment and placement program. Prepared for the Tennessee Board of Regents.

The College Board, (1999b). Placement validity study for the Baltimore City Community College.

The College Board, (2000a). ACCUPLACER Online Users Manual.

The College Board, (2000b). Validity study prepared for Winston-Salen State University, Winston-Salem, NC.

The College Board, (2000C). Analysis of recommended ACCUPLACER cut scores. Prepared for National Louis University.

The College Board, (2003). ACCUPLACER Online Technical Manual. Examination Board, Author.

Attachment 1: Figures

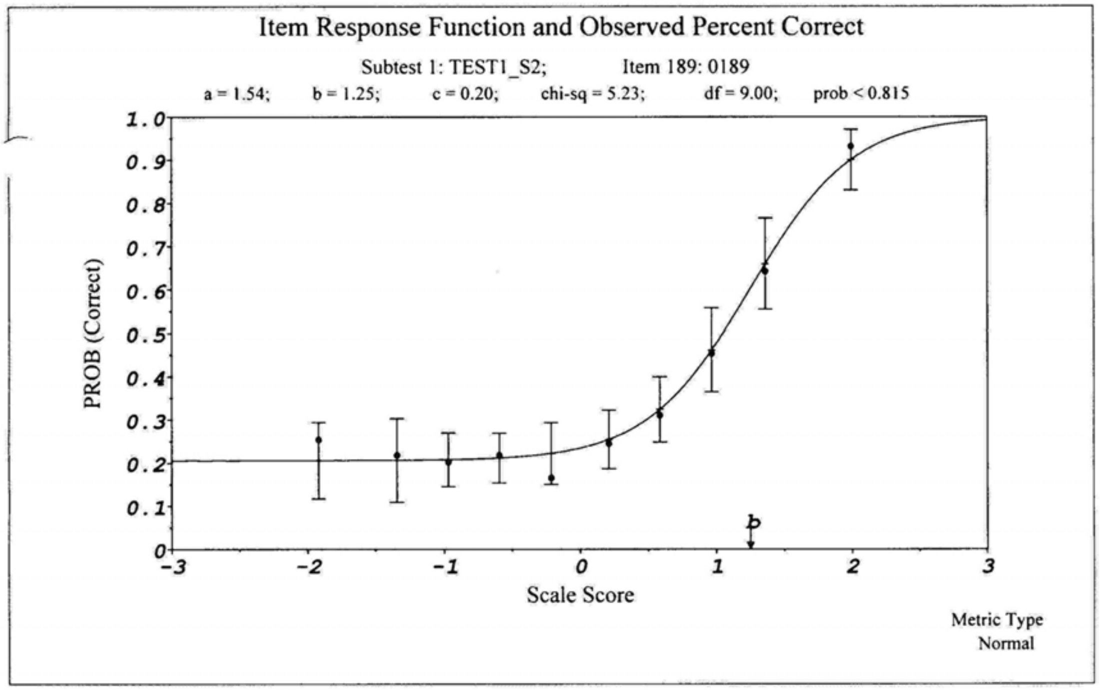

Figure 1: ICC and Item Fit of One Typical Pretest Item

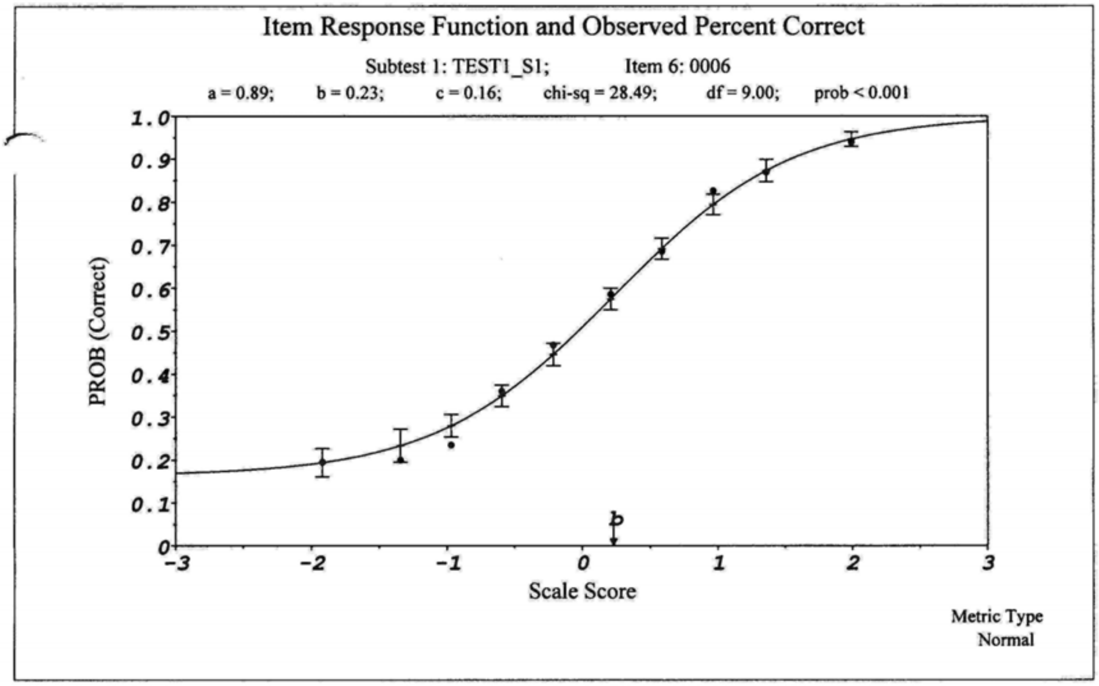

Figure 2: ICC and Item Fit of One Well-Fit Operational Item

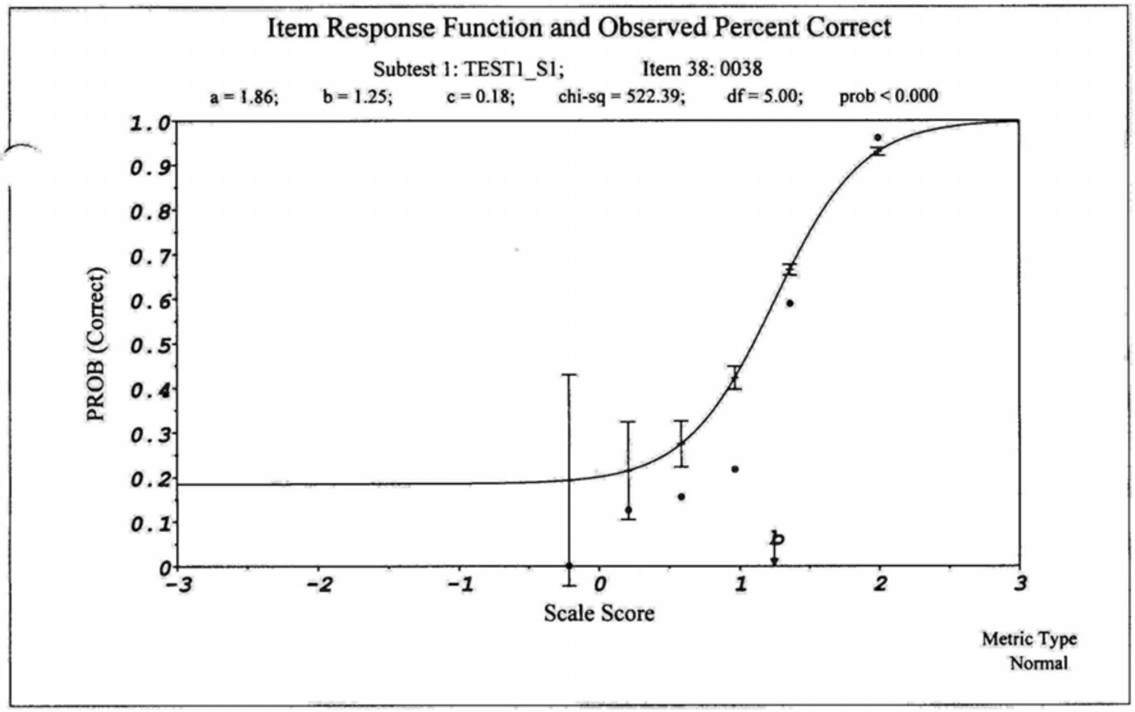

Figure 3: ICC and Item Fit of One Poorly -Fit Operational Item
Calibration Examinee Size=11,609

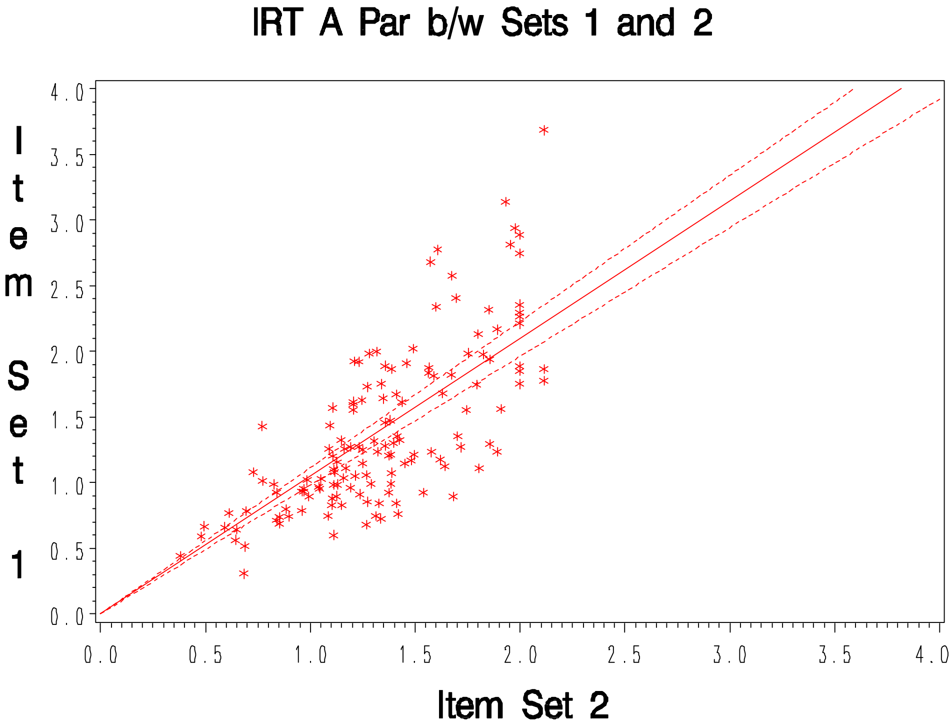

Figure 4: Scatter Plot of *a* Parameter Estimates Between BILOG and LOGIST
Item Set 1=BILOG Estimates, Item Set 2 = LOGIST Estimates

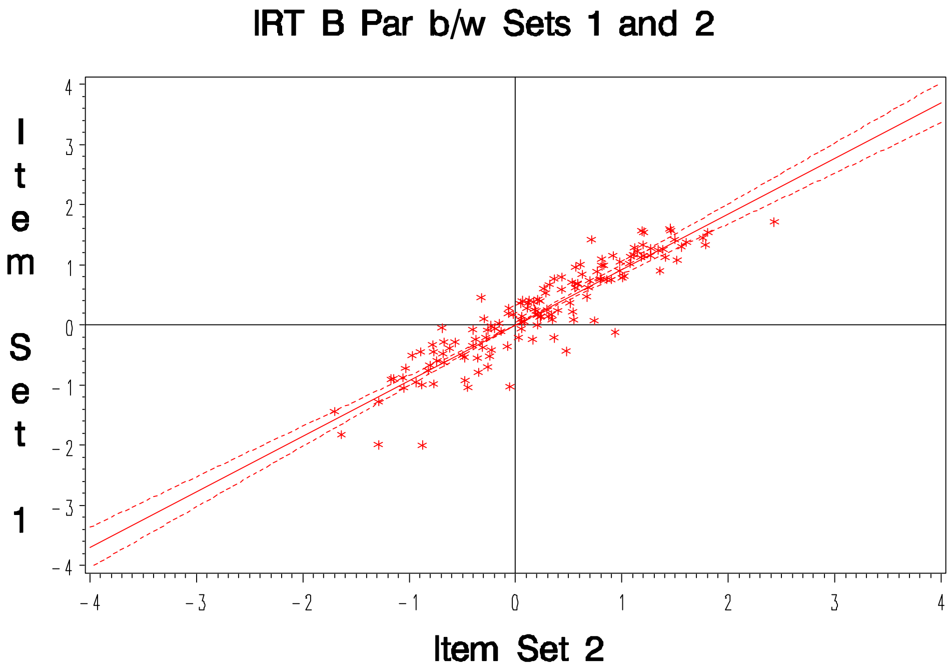

Figure 5: Scatter Plot of *b* Parameter Estimates Between BILOG and LOGIST
Item Set 1=BILOG Estimates, Item Set 2 = LOGIST Estimates

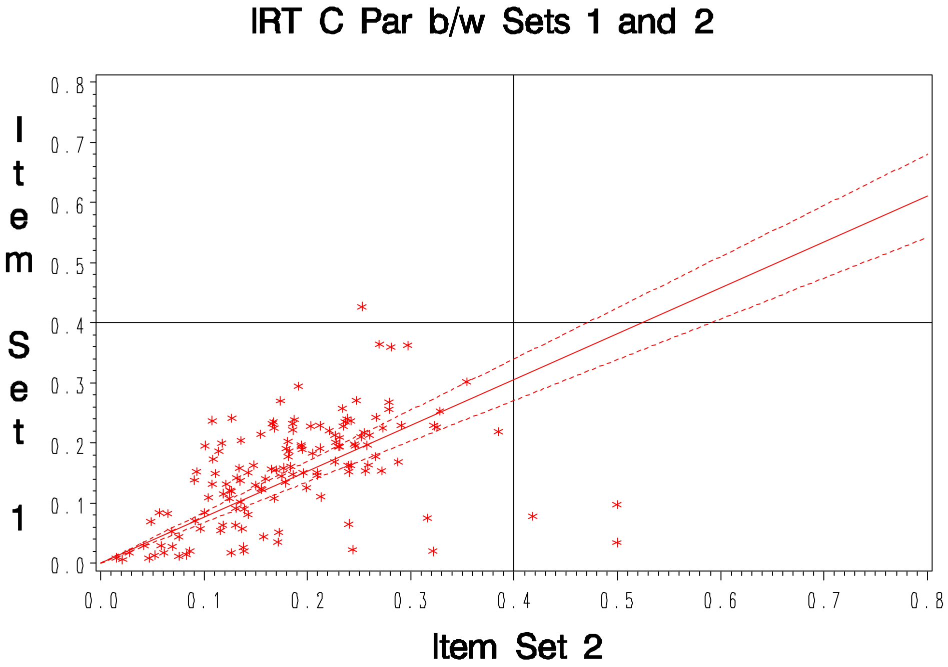

Figure 6: Scatter Plot of *c* Parameter Estimates Between BILOG and LOGIST
Item Set 1=BILOG Estimates, Item Set 2 = LOGIST Estimates

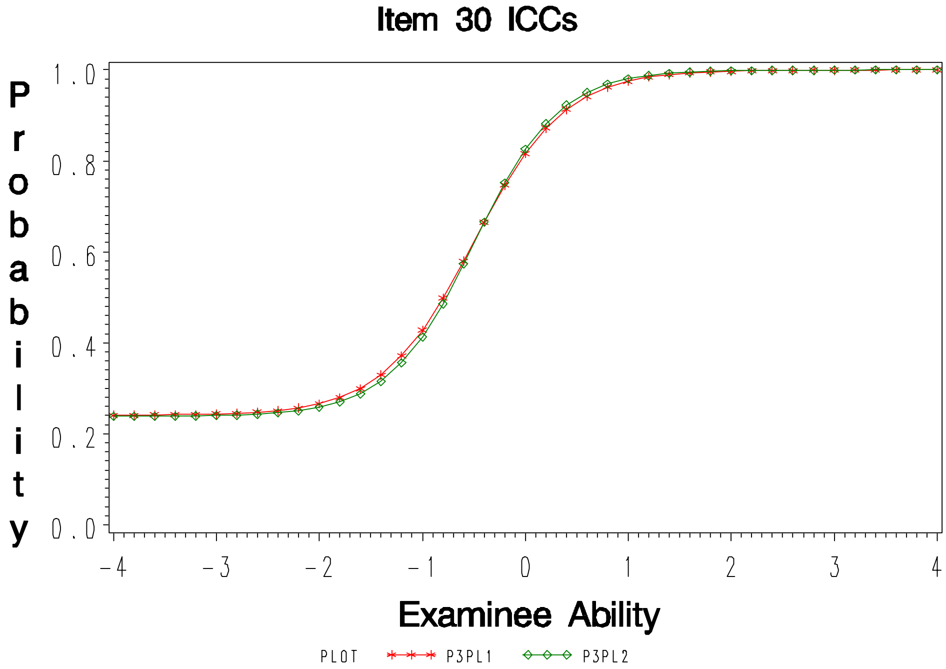

Figure 7: An Item Whose ICCs were Virtually Identical
Between BILOG and LOGIST
P3PL1 = BILOG, P3PL2 = LOGIST

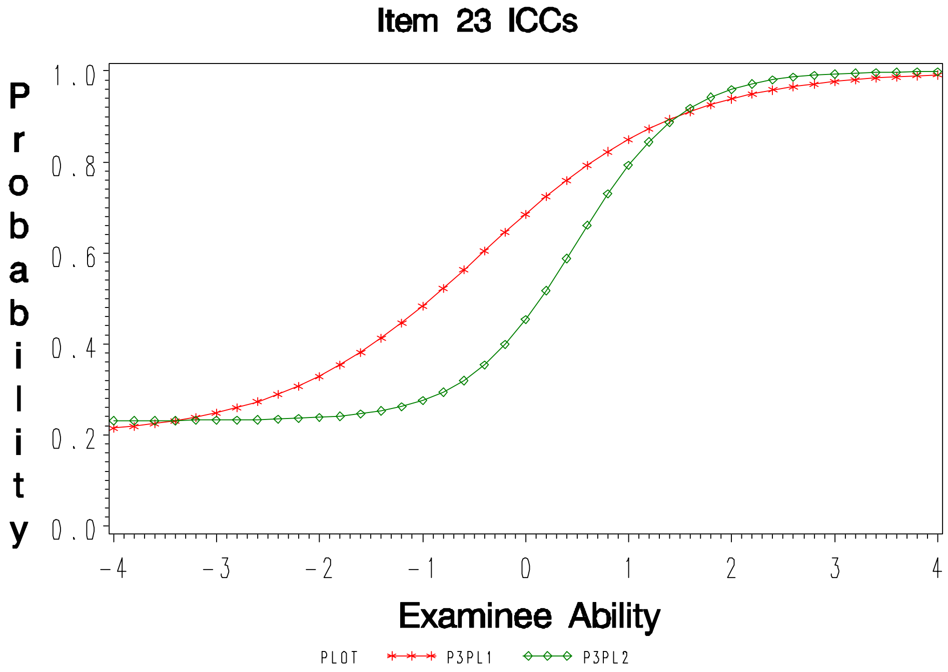

Figure 8: An Item Whose ICCs were Substantially Different
Between BILOG and LOGIST
P3PL1 = BILOG, P3PL2 = LOGIST

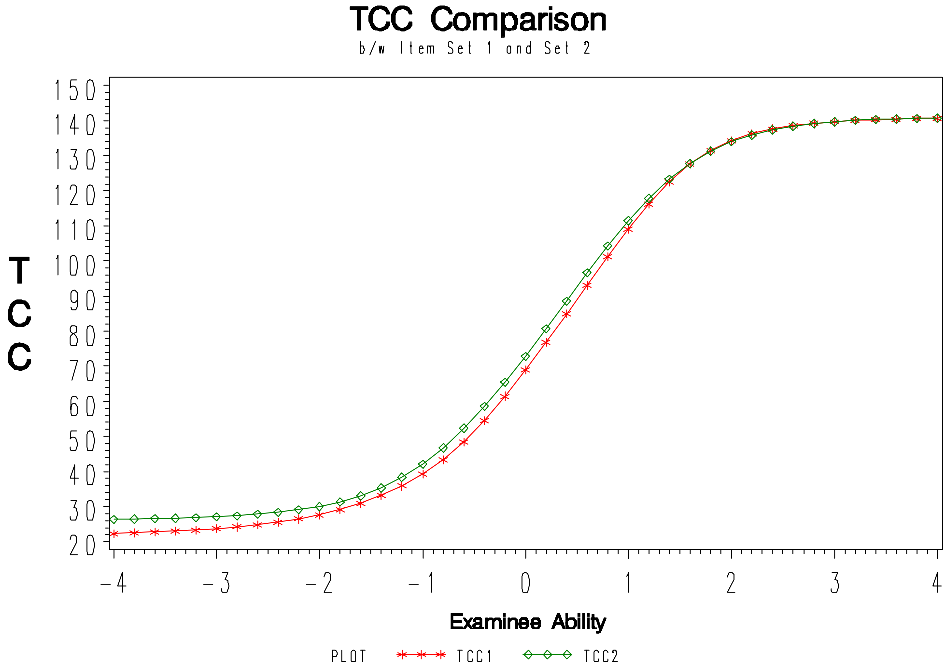

Figure 9: Comparison of Test Characteristic Curves
Between BILOG and LOGIST
TCC1 = BILOG, TCC2 = LOGIST

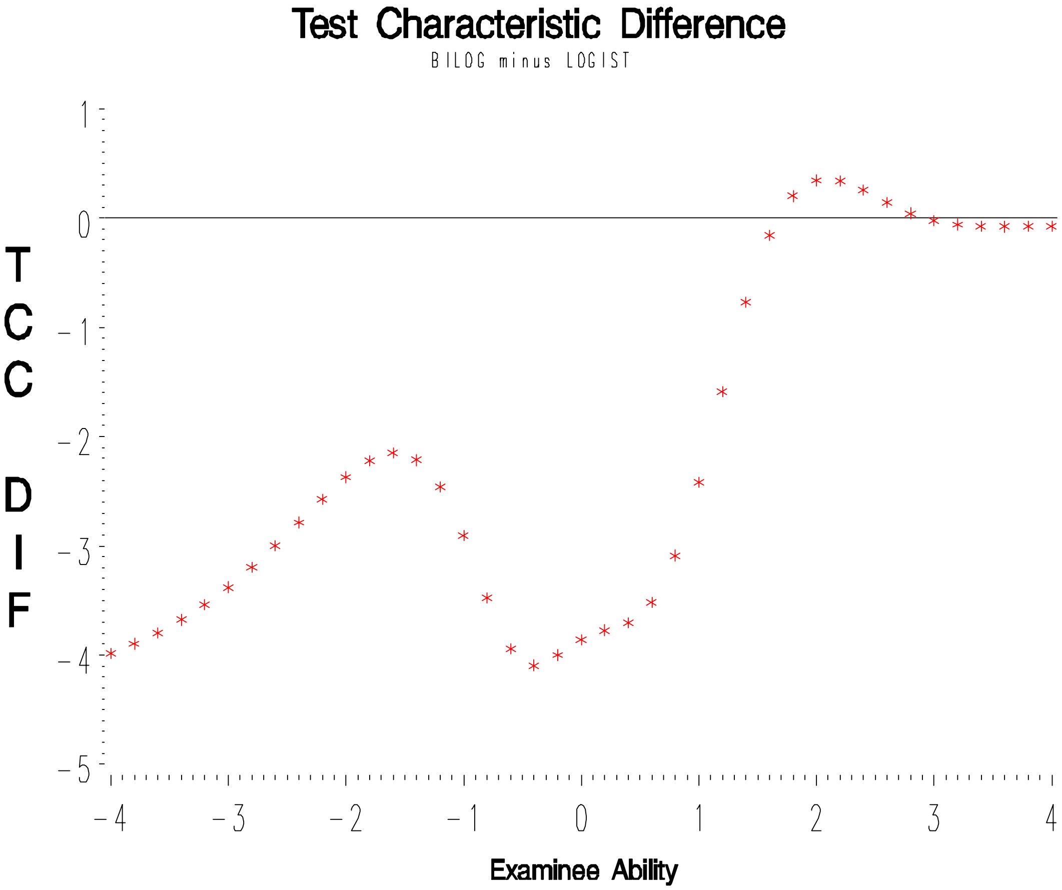

Figure 10: Differences in Test Characteristic Curves
BILOG minus LOGIST

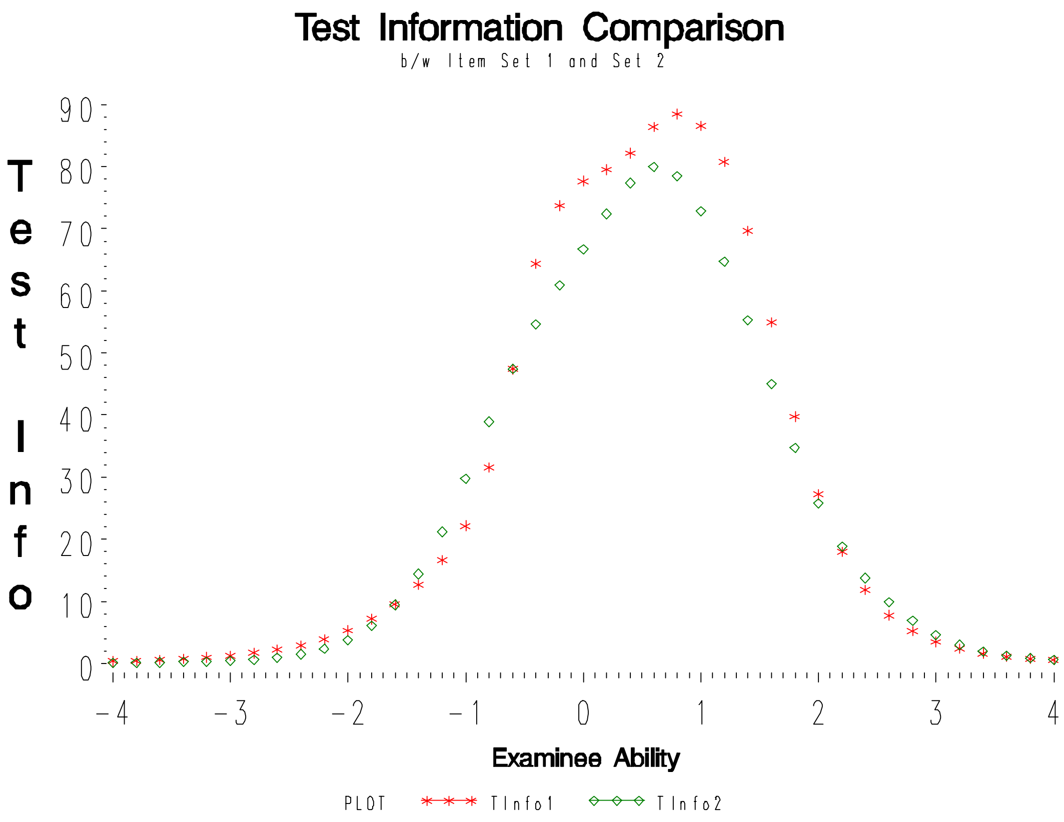

Figure 11: Comparison of Test Information Curves
Between BILOG and LOGIST
Tinfo1 = BILOG, Tinfo2 = LOGIST

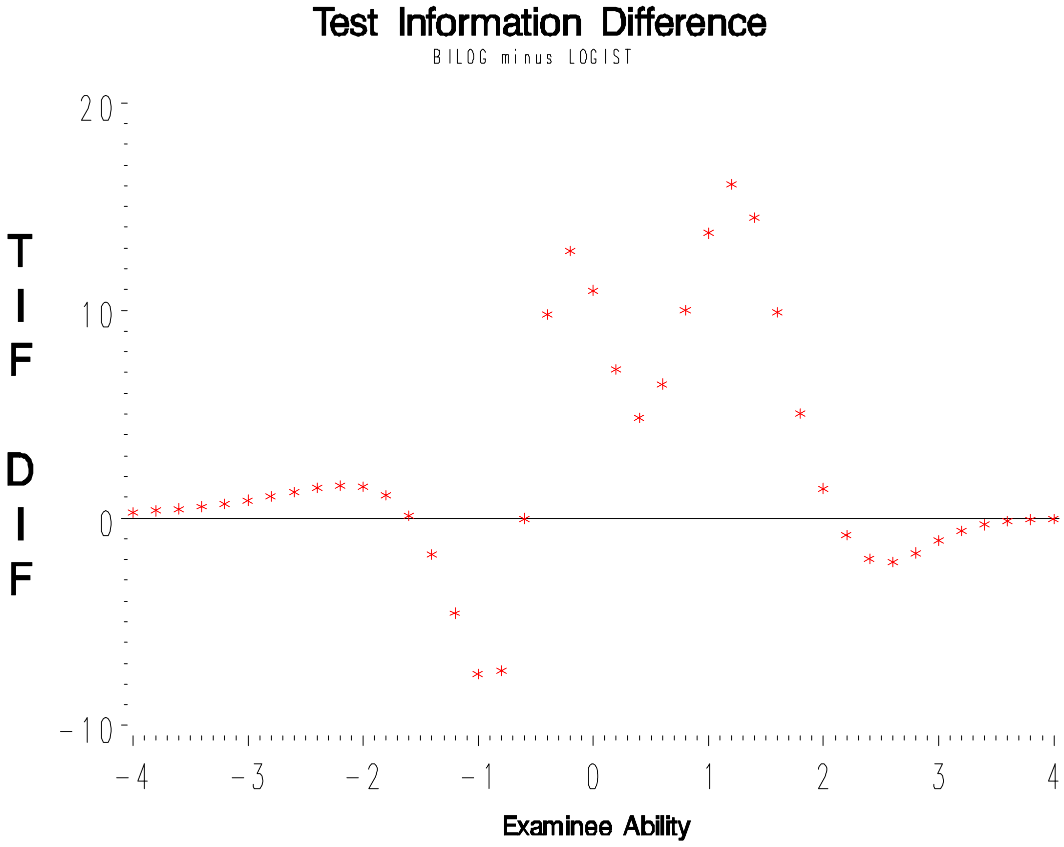

Figure 12: Differences in Test Information Curves
BILOG minus LOGIST

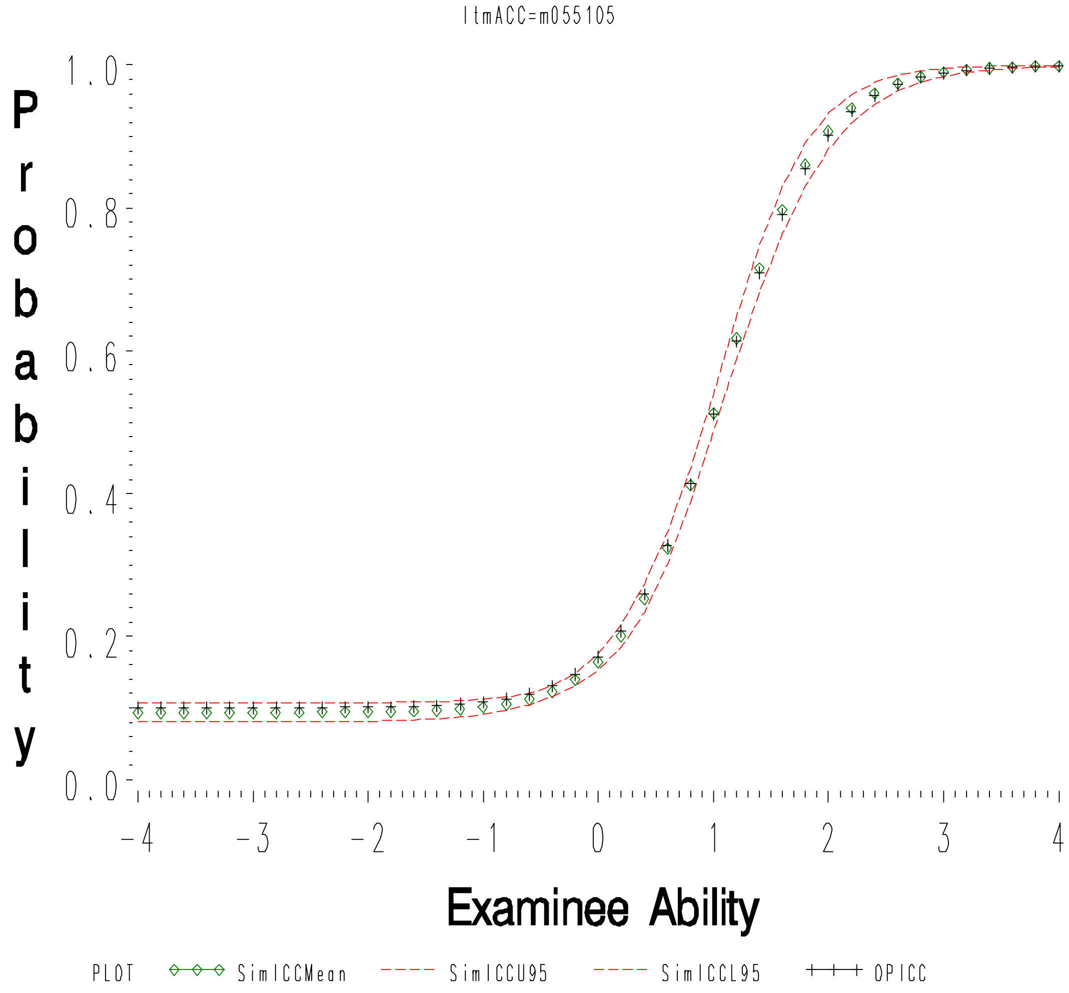

Figure 13: An Item with Target ICC, Mean ICC and Narrow 95% Confidence Interval Bands

SimICCMean = Average ICC across 100 Simulations
SimICCU95 = Upper Bound of 95% ICC Confidence Interval across 100 Simulations
SimICCL95 = Lower Bound of 95% ICC Confidence Interval across 100 Simulations
OPICC = ICC of the Operational Item Calibrated via LOGIST Whose Item Parameter Estimates were Treated as True Parameters for Simulation

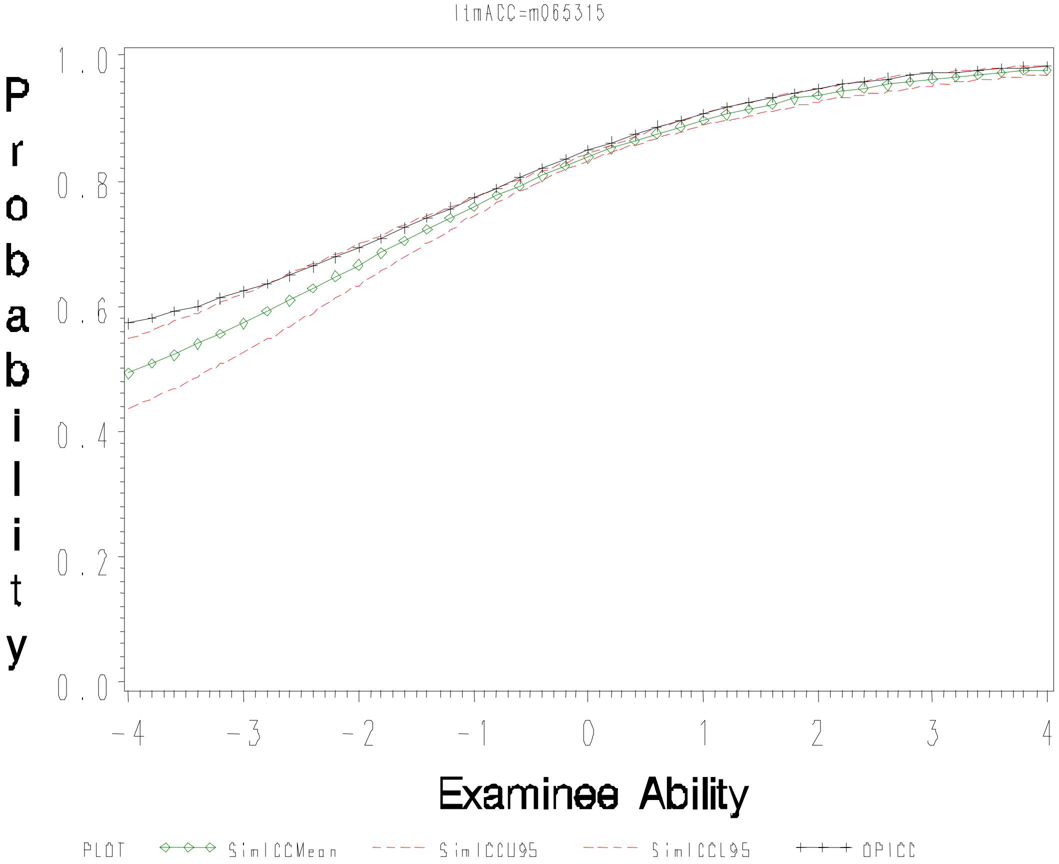

Figure 14: Example for an Item with Target ICC, Mean ICC and Widest 95% Confidence Interval Bands

SimICCMean = Average ICC across 100 Simulations
SimICCU95 = Upper Bound of 95% ICC Confidence Interval across 100 Simulations
SimICCL95 = Lower Bound of 95% ICC Confidence Interval across 100 Simulations
OPICC = ICC of the Operational Item Calibrated via LOGIST Whose Item Parameter Estimates were Treated as True Parameters for Simulation

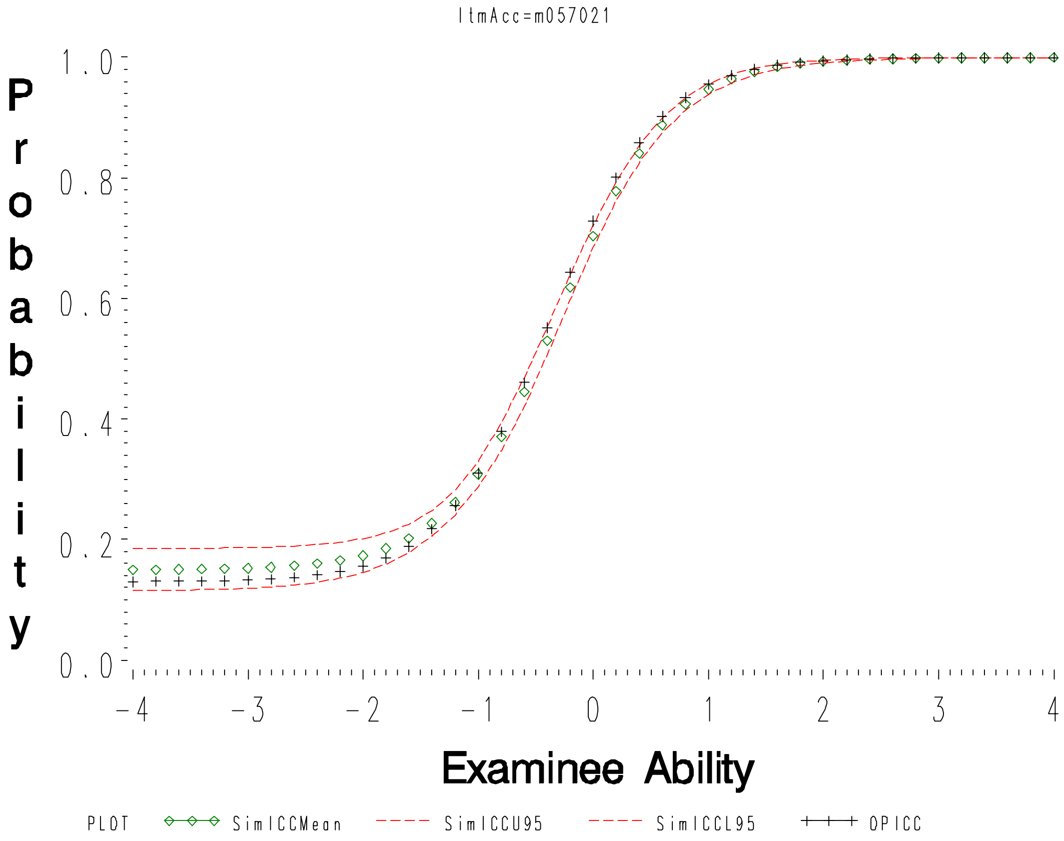

Figure 15: Example for An Item whose BILOG ICC Falls within the 95% ICC Confidence Interval From Simulations

SimICCMean = Average ICC across 100 Simulations
SimICCU95 = Upper Bound of 95% ICC Confidence Interval across 100 Simulations
SimICCL95 = Lower Bound of 95% ICC Confidence Interval across 100 Simulations
OPICC = ICC of an Operational Item Calibrated via BILOG under CAT Data Condition

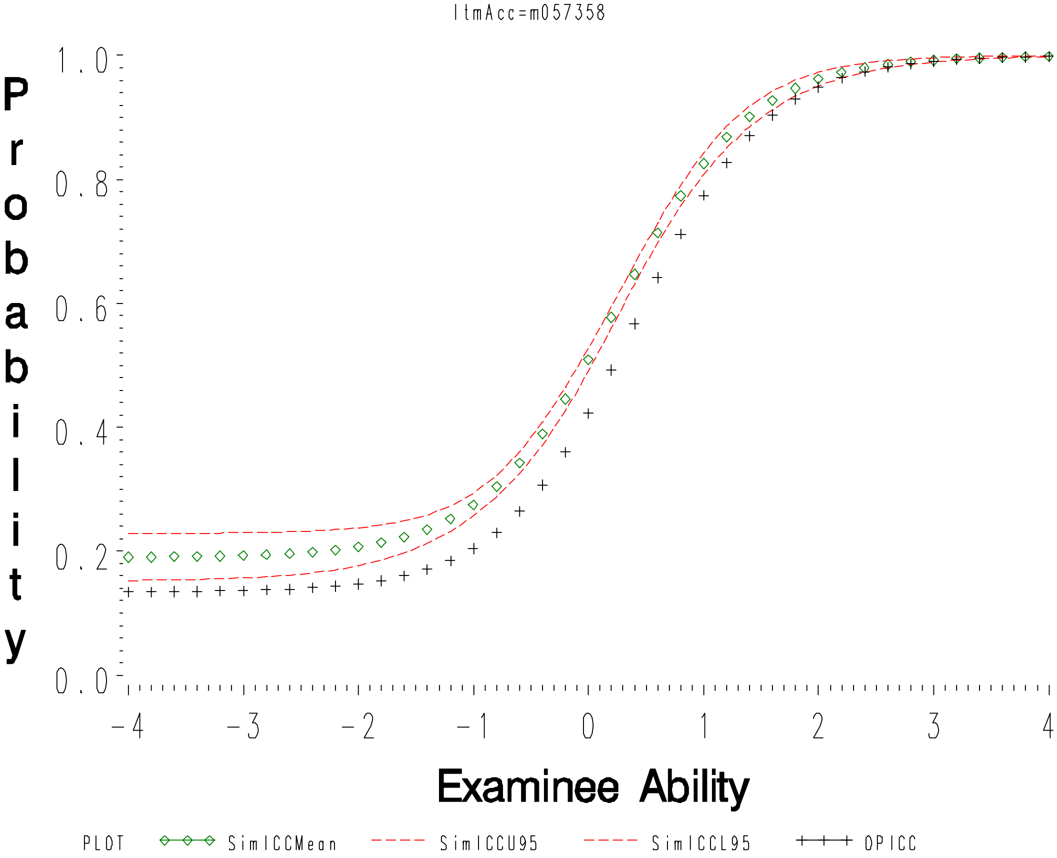

Figure 16: Example for An Item whose BILOG ICC Approximates the 95% ICC Confidence Interval From Simulations

SimICCMean = Average ICC across 100 Simulations
SimICCU95 = Upper Bound of 95% ICC Confidence Interval across 100 Simulations
SimICCL95 = Lower Bound of 95% ICC Confidence Interval across 100 Simulations
OPICC = ICC of an Operational Item Calibrated via BILOG under CAT Data Condition

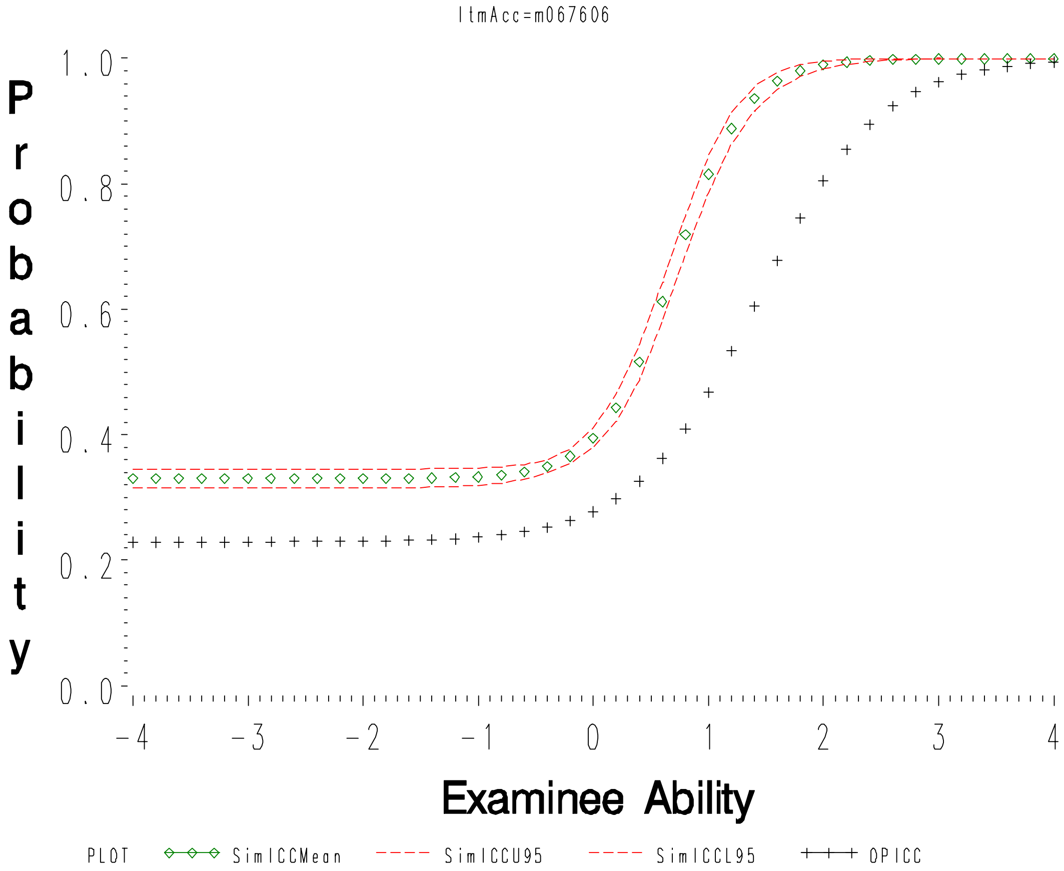

Figure 17: Example for An Item whose BILOG ICC Substantially Deviates from the 95% ICC Confidence Interval From Simulations

SimICCMean = Average ICC across 100 Simulations
SimICCU95 = Upper Bound of 95% ICC Confidence Interval across 100 Simulations
SimICCL95 = Lower Bound of 95% ICC Confidence Interval across 100 Simulations
OPICC = ICC of an Operational Item Calibrated via BILOG under CAT Data Condition