DOCUMENT RESUME

ED 281 864 TM 870 248

AUTHOR	Stocking, Martha L.; Eignor, Daniel R.
TITLE	The Impact of Different Ability Distributions on IRT Preequating.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-86-49
PUB DATE	Dec 86
NOTE	91p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Computer Simulation; Equated Scores; Error of Measurement; *Estimation (Mathematics); *Item Analysis; Latent Trait Theory; Mathematics Tests; *Multidimensional Scaling; Test Items
IDENTIFIERS	*Item Parameters; LOGIST Computer Program; *Pre Equating (Tests); Scholastic Aptitude Test; Three Parameter Model

ABSTRACT

In item response theory (IRT), preequating depends upon item parameter estimate invariance. Three separate simulations, all using the unidimensional three-parameter logistic item response model, were conducted to study the impact of the following variables on preequating: (1) mean differences in ability; (2) multidimensionality in the data; and (3) a combination of mean differences in ability and multidimensionality. One of the Scholastic Aptitude Test mathematical forms (3ASA3) which provided the least acceptable preequating was selected to define true item and person parameters for these simulations. A random sample of 2,744 examinees was used. The LOGIST computer program was chosen to estimate item parameters for the 60 items in 3ASA3 and the 24 items in the equating section fn. Results showed that differences in mean true ability can cause differences in the precision with which a particular estimation procedure estimates parameters. The introduction of a particular kind of multidimensionality in the data can have a large impact on estimation precision when the IRT model is unidimensional. The combination of a slight decrease in mean ability and a particular type of multidimensionality in the data also has a large impact on estimation precision when the IRT model is unidimensional, although the impact is lessened somewhat. (JAZ)

RR-86-49

RESEARCH REPORT

THE IMPACT OF DIFFERENT ABILITY DISTRIBUTIONS ON IRT PREEQUATING

Martha L. Stocking
and
Daniel R. Eignor

Educational Testing Service
Princeton, New Jersey
December 1986

The Impact of Different Ability Distributions on IRT Preequating[1,2,3]

Martha L. Stocking
and
Daniel R. Eignor

Educational Testing Service

October 1986

[1] An earlier version of this paper was presented at the annual meeting of AERA, San Francisco, 1986.

[2] This study was supported by Educational Testing Service through Program Research Planning Council funding.

[3] The authors would like to acknowledge the advice of Marilyn Wingersky and the assistance of Nancy Wright in performing this study.

ABSTRACT

Item response theory preequating depends upon item parameter estimate invariance. The impact of differences in true ability on the invariance properties of item parameter estimates was studied with simulated data. Using real SAT-mathematical data that had produced unsatisfactory preequating results to suggest hypotheses, three explanatory models were investigated: 1) differences in mean true ability, 2) a certain type of multidimensionality, and 3) a combination of differences in mean true ability and multidimensionality. This latter model produced results consistent with the real data.

The Impact of Different Ability Distributions on IRT Preequating

Martha L. Stocking
and
Daniel R. Eignor

INTRODUCTION

In item response theory (IRT), when model assumptions are satisfied, true item parameters do not change even when considered across samples with different true abilities from the sane population. Likewise, true abilities do not change, even when considered in reference to different sets of items (Lord, 1980). This is called the 'invariance' property of the true item and person parameters.

The invariance property of true item parameters suggests that it is possible to equate a test before it is actually administered, as long as true item parameters are known. This is called 'preequating'. The invariance property of true abilities suggests that adaptive testing, where individuals take different sets of items, is possible.

How well either of these two novel ideas works in practice depends not upon the true item parameters or person parameters, but rather, on ESTIMATES of them. To the extent that estimates fail to approximate truth, both preequating and adaptive testing will fail. While there may be many specific reasons why estimates do not approximate truth very well, the reasons can generally be classified into two broad categories: reasons having to do with the imprecision introduced by various estimation procedures currently in use; and reasons having to do with the failure of the data to satisfy the underlying assumption(s) of the particular IRT model used.

Two recent studies of preequating SAT verbal and mathematical data using the three-parameter logistic (3PL) item response model showed disappointing results in the face of reasonable evaluative criteria (Eignor, 1985; Eignor & Stocking, 1986). These large scale studies showed that only one of two verbal preequatings was adequate, and that neither of two mathematical preequatings was adequate. Explorations of many reasonable explanatory hypotheses were conducted, but no definitive answers were found. It was suggested that differences in abilities across samples might somehow cause the results found, namely that the tests under study had higher raw score to scale conversions, i.e., appeared to be more difficult, when preequated than when equated using intact final form data from a regular administration. This hypothesis was further strengthened by two observations:

1) Items tended to be harder in pretest form than in intact operational administrations, i. e., the b 's were higher.
2) Pretest samples tended to have lower abilities than intact form administration samples, as measured by the scaled

score means the pretest samples attained on the intact forms accompanying the pretests.

Sample differences could cause the results found in the preequating studies either because such differences introduce different estimation errors, or because they in fact represent a violation of model assumptions, or both.

This current study attempts to simplify the study of preequating by using simulated data. Because the data are simulated, one can more easily study the effects of sample variation on preequating. Three separate simulations, all using the unidimensional three-parameter logistic item response model, were conducted. These simul~~t~~ations were designed to study the following variables:

1) Mean differences in ability.

   Samples of data that vary only by a shift in the average true ability were simulated. While different estimation errors do impact preequating, the results of this simulation did not explain the previous real-data results.

2) Multidimensionality in t e data.

   A certain type of multidimensionality was introduced into different simulated samples. The data were analyzed with a unidimensional item response model, thus violating model assumptions. The effects on preequating were partially consistent with results from real data, although much larger.

3) Mean differences in ability and multidimensionality in the data.

This simulation combined the two types of sample variations studied above. The effects on preequating were more moderate than those found in the second study, although larger than those seen with real data. These results, however, were completely consistent with the real data results.

METHODOLOGY

The Definition of Truth

One of the two SAT mathematical forms from the previous studies was selected to define true item and person parameters for these simulations. The form chosen, 3ASA3, provided the least acceptable mathematical preequating. Using a random sample of 2744 examinees from the operational administration of this form with equating section fn, item parameters for the 60 items in 3ASA3 and the 24 items in fn were estimated using LOGIST (Wingersky, Barton, & Lord, 1982). These item and person parameter estimates were then used as realistic true item and person parameters. Table 1 gives summary statistics for these true parameters.

Insert Table 1 about here

When 3ASA3 was first administered as an intact test form, it was equated to the familiar College Board scale for score reporting purposes. The particular equating chosen at that time was a linear one. For purposes of these simulations, this linear equating will be, by definition, the 'true' equating associated with the true item and person parameters.

Using the frequency distribution of observed scores for all individuals who took 3ASA3 at this first administration, a 'true' scaled score mean of 485, and a 'true' scaled score standard deviation of 113 were computed. Comparisons among simulated equatings will frequently be made in reference to these 'true' values.

First Simulation: Mean Differences in Ability

The first simulation was designed to explore the hypothesis that preequatings that produce higher scaled score means (meaning that the preequated test appears to be more difficult) result from less able preequating samples. While this idea is plausible, it challenges the efforts to produce item parameter estimates that exhibit the invariance property of true item parameters.

The simulation was designed to mirror, as much as possible, differences observed in the summary statistics for real data. For 3ASA3, 13 out of the 14 samples on which items were pretested had lower scaled score means on the intact forms administered with the pretests than the sample taking 3ASA3 when given as an intact form. The lowest scaled score mean on all intact test forms given with 3ASA3 items being pretested was 441. The scaled score mean for 3ASA3 when given in intact form was 485. Using results from a typical IRT equating, this 44 scaled score point difference translates into a difference of about .35 on the IRT ability metric. Sample scaled score standard deviations varied only from 110 to 117 in the previous studies. Therefore, no attempt was made here to simulate differences in variances among the simulated samples.

Simulated Samples

Using the true abilities for 3ASA3 and fn, four different distributions of true abilities were independently generated with progressively lower true ability means (0, -.35, -.70, -1.05). These particular levels were chosen for two reasons: 1) the difference between the first two (.35) matches the largest mean decrease found in the real data, and 2) it was hoped that futher decreases would result in exaggeraged and therefore easily detectable effects resembling real-data results. Samples of N = 2500 simulees were then drawn from each distribution. The bottom portion of Table 2 presents the results of this sample selection.

Insert Table 2 about here

Responses to each item in 3ASA3 and fn were then generated using the true abilities for each sample and the true item parameters for all 84 items.

Estimation of Item Parameters and Abilities

LOGIST (Wingersky, Barton, & Lord, 1982) was used to calibrate all items and abilities in a single concurrent execution, with equating items fn used as an anchor to set the scale. This method, described in detail in Petersen, Cook, and Stocking (1983), has provided satisfactory parameter scaling results in a number of studies. The N = 10000 and n = 264 data matrix can be represented as follows:

Items People	fn	3ASA3-1	3ASA3-2	3ASA3-3	3ASA3-4
Sample 1	x	x			
Sample 2	x		x		
Sample 3	x			x	
Sample 4	x				x

In this matrix, an x indicates that a group of items is taken by a particular group of examinees; a blank indicates that group of items is not administered to a group of examinees. The above design produces four different sets of item parameter estimates for the 3ASA3 items. Each set of estimates differs only in the mean ability level of the group used for estimation. This mirrors a calibration of 'pretest' items (taken by samples 2, 3, and 4) and, simultaneously, a calibration of 'operational' items taken by sample 1.

Scaling of Estimates

LOGIST establishes the metric upon which parameter estimates are reported by setting the mean and standard deviation of a truncated distribution of ability estimates to zero and one, respectively. The true item and ability parameters are on a different scale. Therefore, before any comparisons can be made between estimated and true parameters, a scaling transformation is required.

Sample 1 comes from the original distribution of true abilities. If Sample 1 estimated abilities differed from the true abilities by only a scaling factor, one could use the relationship between these estimated abilities and true abilities to determine the appropriate scaling transformation. Since the estimates contain errors, one can approximate the scaling transformation by determining for Sample 1 the transformation necessary to make robust measures of location and scale of the estimated abilities equal to robust measures of location and scale of the true abilities. This linear transformation can then be applied to all estimates in the LOGIST run to place them on the same scale as the true values.

Comparison of Estimated and True Parameters

Summary statistics for the estimates of item and person parameters after the scaling transformation are presented in Table 2. In this table, it can be seen that the mean true ability as well as the mean estimated ability decrease across the four samples, a consequence of the study design. It is important to remember, however, that during the calibration process, parameters for all four samples were estimated simultaneously. LOGIST standardizes its results using the mean and standard deviation of *all* estimated abilities. This mean will lie somewhere between the means for Samples 2 and 3. Therefore, Samples 2 and 3 are closer to the overall mean true ability during the calibration, and Samples 1 and 4 lie further away.

Simple "box and whisker" plots that graphically show the relationships among the distributions of estimated abilities are given in the top part of Figure 1. The horizontal axis in this figure represents ability. The left and right asterisks mark the 10th and 90th percentiles of the distribution. The left side and right sides of the box mark the 25th and 75th percentiles. The vertical bar in the box interior marks the 50th percentile.

Insert Figure 1 about here

Figures 2 through 5 compare estimated item parameters and estimated abilities for test 'forms' 3ASA3-1, 3ASA3-2, 3ASA3-3, and 3ASA3-4 with the true values. The different symbols on a single plot indicate the behavior of item parameters shown in other plots. Examination of these figures leads to the following observations:

Insert Figures 2, 3, 4, and 5 about here

1) Estimates of item discriminations using Sample 1 data are generally too low. Sample 1 was the most able sample. Estimates of item discriminations from Sample 2 and 3 data are reasonably good. Estimates of item discriminations from Sample 4 data are generally too high. Sample 4 was the least able sample.

2) The item difficulties for the samples closer to the overall mean true ability (2 and 3) are slightly better estimated (have less scatter) than item difficulties from the more extreme samples (1 and 4).

3) The difficulties for easy and hard items are less well estimated than those for less extreme items, regardless of the sample used. The most able sample (Sample 1) has more overestimated hard items. The least able sample (Sample 4) has more overestimated easy items.

4) The less able the sample, the better the estimates of c become.

5) Low and high abilities tend to be overestimated. Because Sample 1 is the most able sample, it has the greatest number of overestimated high abilities. Because Sample 4 is the least able sample, it has the greatest number of overestimated low abilities. The two middle samples have fewer overestimated abilities than the two extreme samples because they are closer to the overall mean true ability.

The fact that the estimation procedure does not recover the true parameter values is not suprising. Any estimation procedure is imperfect. But it is important to understand why the procedure is systematically imperfect, because this will explain how estimation errors impact subsequent equatings. In this case, the explanation proceeds as follows:

1) It has been obeserved that extreme abilities (either high or low) can be overestimated when the item parameters are not known (Lord, 1975, p. 10). While an explanation for this phenomena is currently under development, it is important to to note that in other simulation studies, different estimation errors have sometimes been noted. The overestimation of high abilities is almost always observed. However, low abilities are sometimes observed

to be underestimated (Wingersky, 1985) as well as overestimated.

In addition, Lord (1975) shows that the overestimation can be greater for low abilities than for high abilities. Examination of the Figures 1 through 4 show that this is the case here also. The extreme samples, 1 and 4, contain more overestimated abilities than the two middle samples. In addition the degree of overestimation for the low abilities in Sample 4, the least able sample, is greater than for the high abilities in Sample 1, the most able sample.

2) In Sample 4, difficulty parameter estimates for easy items tend to more overestimated than parameters estimated for difficult items because lower abilities are more overestimated than high ones. In Sample 1, hard items tend to be more overestimated than easy ones because of the overestimation of high abilities. This is so because the overestimated abilities give erroneous information about item location.

3) Wingersky and Lord (1984) show that there is a positive sampling correlation between estimat·.s of a and b when the item is easy, and a negative sampling correlation when the item is difficult. For Sample 4, the least able sample, all but one of the estimated a 's is too high for

easy items (b < -1.0). For Sample 1, the most able sample, all but one of the estimated a 's is too low for hard items (b > +1.0).

To summarize: estimation errors found for extreme abilities are reflected in estimation errors for item difficulties. Because of the sampling correlations between item difficulty estimates and discrimination estimates, predictable estimation errors then occur for the item discriminations.

Equating Results

Of primary importance in this study is the analysis of equating results when item sets have been calibrated on samples of different ability. Figure 6 shows the results of IRT equatings of forms 3ASA3-2, 3ASA3-3, and 3ASA3-4 to form 3ASA3-1. The figure displays both the equating and equating residuals plots. The linear criterion equating is the 'true' equating of the intact form 3ASA3 to the 200 to 800 score metric.

Referring to the residual plots, it may be seen that for small differences in mean true ability for the calibration group, the impact on equating is really quite small, less than 5 scaled score points at all levels of raw scores. For the largest difference in mean true ability, the impact is greater. For higher raw scores, it can be as much as 15 scaled score points.

Insert Figure 6 about here

Using the frequency distribution of scaled scores obtained when 'true' form 3ASA3 was operationally equated, the top part of Table 3 summarizes the equating results in terms of the scaled score means and standard deviations. From these numbers, it may be seen that small sample differences cause about a one-point difference in scaled score means. The largest sample difference, from the least able sample, is about 5 scaled score points.

Insert Table 3 about here

How do these equating results compare with the real-data preequating results? The differences are striking:

1) Differences in mean true ability of 1/3 to 2/3's of a standard deviation have only a very slight impact on equating. The magnitude of the differences for real-data results was even larger than the differences seen for the least able sample. This sample had mean true ability about one standard deviation below the most able sample. However, the real data contained no sample differences this large. Hence, this simulation cannot explain the real-data results.

2) The direction of the equating differences is exactly the opposite in the simulation from that found in real data. Here, we find that 3ASA3-4, calibrated on the least able sample, appears easier than it should. In real data, the preequating indicated a harder test, not an easier one.

We can explain the equating differences found for our simulated data at least partially in terms of the item parameter estimation errors previously described. It is clear that the difference in item parameter estimation errors for Samples 1, 2, and 3 have only a small impact on equating results. The impact on equating begins to become important only for the least able sample, Sample 4.

Figure 7 compares the item parameter estimates from the least able sample with those from the most able sample, Sample 1. In these plots, different plotting symbols in one plot indicate the behavior of the parameter estimates in another plot. Estimates of the a 's for Sample 4 are higher than estimates for Sample 1. This is true since the a 's were underestimated from Sample 1, and overestimated from Sample 4. The mean estimated a for Sample 4 is 1.05, while that for Sample 1 is .95. The estimates of item difficulty are not that different; the Sample 4 mean is -.01, while the Sample 1 mean is +.01.

Insert Figure 7 about here

The resulting impact on equating is most easily seen in Figure 8, which plots the test characteristic curves for all four simulated forms. Because of the overestimation of the a 's in the least able sample, the test characteristic curve for 3ASA3-4 is shifted to the left of the others. For any value of true ability above .5, the number right true score will be higher on this form than on the other forms. Hence, form 3ASA3-4 appears easier for individuals of moderately high true ability than the other forms. There is little difference among the test characteristic curves at middle and low ability levels.

Insert Figure 8 about here

Thus, one can explain, at least partially, the differences found in the simulated equatings through differences in parameter estimation errors caused by different samples of true ability. Unfortunately, this does not illuminate the real-data results from the previous studies.

How Big Is Bad?

A separate aspect of equating differences can be explored using data from this simulation. It was previously observed that scaled score mean differences of up to 5 points can result from different samples. Scaled score mean differences in the real-data study were up to 13 points. While smaller differences are better, how can one understand the importance of these differences?

One method of evaluating differences is to compare equatings where one set of item parameters is estimated and the other set of item parameters are considered to be the truth. Figures 9 shows equating results when forms 3ASA3-1, 3ASA3-2, 3ASA3-3, and 3ASA3-4 are equated to 'true' test form 3ASA3. The differences are quite large when compared to the corresponding equatings when item pa~ameters for both forms are estimated. The differences seen for form 3ASA3-1 indicate the magnitude that can be expected on the basis of what is predominately estimation error, since this form was taken by Sample 1, whose mean true ability was the same as the definition of truth. Other equating differences result from a combination of estimation error alone and estimation error due to differences in abilities.

Insert Figure 9 about here

The results are summarized in terms of scaled score means and standard deviations in the middle portion of Table 3. As a result of only estimation errors, a difference in mean scaled scores of about 2 scaled score points is observed. Equating errors from differences in estimation errors resulting from differences in true ability can be higher, about 3 scaled score points.

It is interesting to note that, although Sample 1 and Sample 4 are about equally as far from the overall true mean ability in the calibration, the type of estimation errors made for these two outlying samples has a very different impact on equating errors. Samples 1, 2, ard 3 have about a 2- to 3-point increase ir. mean scaled score over true mean scaled score; Sample 4 has about a 3-point decrease in mean scaled score over true mean scaled score.

Conclusions from the First Simulation

Differences in mean true abilities can cause differences in equatings. For small differences in mean true abilities, these equatings differ by about what one would expect on the basis of estimation errors alone. For a large difference in true ability, the difference in equated means is about twice that. These equating differences are at least partially explainable on the basis of the known magnitude and direction of estimation errors when samples differ in mean true ability. However, the direction of equating errors is opposite to that found in the previous studies with real data.

Second Simulation: Multidimensionality in the Data

From the results of the first simulation, it is clear that the explanation of poor preequating results found with real data does not lie solely with the imprecision of the estimation procedure. This second study was designed to explore th. other category of potential problems in parameter estimation: the failure of the data to satisfy the underlying assumption(s) of the particular IRT model used.

The Eignor and Stocking (1986) results were reexamined, this time in terms of the abilities estimated by LOGIST for every sample of examinees that contributed to the calibration of pretest items. Table 4 shows the summary statistics for the real data used to preequate test form 3ASA3. Each sample is labeled, and the number of pretest items contributed by this sample is in parentheses by the sample designation. Samples are listed in decreasing order by median estimated ability. Percentile information is also displayed graphically in "box and whisker" plots in Figure 10.

Insert Table 4 and Figure 10 about here

The use of these distributions to make inferences about distributions of true abilities is not strictly correct, since estimated abilities have different properties than true abilities. In addition, the number of items on which an ability estimate is based differs by sample; hence, estimation errors will be different for each sample. However, a number of observations can be made.

There are four pretest samples that contribute over half of the pretest items that appear in 3ASA3. They are designated C1613, C1614, C2314, and C2318. The mean estimated ability for these samples is about .2 to .4 standard deviations below the mean estimated ability of the operational sample (3ASA3-Oper.). The standard deviations of estimated ability vary by at most .05. It is this kind of mean shift with no change in variance that the first simulation was designed to study. It can be seen from the results of the first simulation that mean differences alone cannot account for the preequating results found with the real data.

Of greater interest in Figure 10 is the comparison of the differences in the percentiles shown. Here one sees that the distributions of estimated abilities are distorted, not merely shifted, when compared to the distribution of estimated abilities for the operational form. For the four pretest samples contributing over half the items, the distributions are shifted lower when compared to the operational form, but the shift is larger at the 25th, 50th, and 75th percentiles than it is at the 10th or 90th percentiles.

These samples are supposed to be samples from the same overall population, although we have no way of proving the truth of this assertion. It is possible, of course, that repeated samplings from the same population can give rise to such distortions. It is also plausible that such distortions can result from some mechanism that makes a unidimensional IRT model inappropriate for these data.

It is not hard to advance hypotheses about circumstances that could introduce multidimensionality. Among the many possible ones are the effects of improved teaching methods on more recent samples of students, changes in emphasis and curriculum that took place between pretest and operational administration, and the ability of examinees to recognize and therefore have different motivation on pretest sections. This latter situation could very well be applicable for the real-data results.

In current SAT administrations, test sections that contain items that are being pretested are labeled in a manner that is indistinguishable from operational sections, and appear in different locations in different test booklets. This has not always been the case. Less than half the items in the final form 3ASA3 were contributed by 12 pretest sections that had labeling that could be distinguished from that of operational sections; these pretest sections have designations beginning with X or Z. More than half the items in the final form 3ASA3 were contributed by 4 pretest sections that had indistinguishable labeling, but were always located in the same positions within the test booklets. Thus there is some reason to believe that any of the prestest sections contributing items to final form 3ASA3 could have been subjected to recognition and, therefore, motivational effects.

In these studies, we focus our attention on the four pretests that were administered in 'fixed' rather than 'variable' positions for three reasons: 1) these prestests contributed over half of the items to final form 3ASA3, 2) these pretests were administered most recently and therefore within the current social climate of 'test wiseness' encouraged by coaching schools, and 3) for students not possessing special information, a pretest section in a fixed position is probably easier to detect than a pretest section having a label based on a distinguishable labeling scheme.

The Multidimensional Model

McDonald (1982) provides a broad framework, based on nonlinear factor analysis, for the classification of unidimensional and multidimensional models. The particular model chosen here falls into McDonald's general category of nonlinear multidimensional models.

In the particular model used in these studies, examinee responses to some items are generated using 3PL item response functions and a certain true ability. Responses to other items for the same examinee are generated using 3PL item response functions and a second true ability. The second true ability is related to the first through a discontinuous step function.

This model in effect forces the 3PL model to hold for all item response functions but assumes that examinees respond to some items with one ability and to others with another. This is different, and therefore less familiar, than the multidimensional linear model often used in IRT multidimensionality studies (see Drasgow & Parsons, 1983, for example) but seems more intuitively appealing in the present circumstances.

Simulated Samples

Using the true abilities, three new samples of N = 2500 each were drawn with no modifications to the true ability distribution. The 60 items in 3ASA3 were considered to be 'operational' form 3ASA3-5. Two nonoverlapping random subsets of items from 3ASA3 were formed, each containing 30 items and designated as 3ASA3-5A and 3ASA3-5B. These two smaller subsets are to be considered as pretest items for equating purposes: each will be administered to different samples, and the resulting item parameter estimates will be combined to constitute a full 60-item test form. Using true parameters, responses were generated for simulees as follows:

----- Sample 1 responded to the 24 items of equating section fn and test form 3ASA3-5.

----- Sample 2 responded to the 24 items of equating section fn and the 30 items in 3ASA3-5A.

----- Sample 3 responded to the 24 items of equating section fn with abilities sampled from the same true ability distribution as the other two samples. However, when responding to the 30 items in 3ASA3-5B, their true abilities were distorted. This was done in the following manner:

1. If true ability was less than or equal to -1, no change in ability was made.
2. If true ability was between -1 and -.5, the simulee responded with an ability equal to true ability minus .2.
3. If true ability was between -.5 and +.5, the simulee responded with an ability equal to true ability minus .4.
4. If true ability was between .5 and 1.5, the simulee responded with an ability equal to true ability minus .6.
5. If true ability was above 1.5, no change in ability was made.

These particular distortions were chosen to reflect distortions that might have caused the results observed for distributions of estimated ability from the real data. There are at least two intuitively appealing rationales that can be used to justify them. The first rationale runs along the following lines: individuals of low ability are not aware of clues that might change their motivation, so their behavior remains the same. As true ability increases, so does sensitivity to such clues and the ability to take advantage of them. Very able individuals, however, have no need to use such clues and continue to perform at the same high level as before.

A second appealing rationale focuses on targeted improvements in teaching and curricula. Individuals of very low ability are not in the targeted group. As true ability increases, the improvements become more appropriate and have a larger impact. Very able individuals, however, have no need of improved teaching or curricula since their ability is so high that they will learn the appropriate material regardless of how poorly or well it is taught.

The results of the generation of samples of true ability are shown in the bottom portion of Table 5. For the third sample, only the distorted abilities used to generate responses to 3ASA3-5B are shown. The true abilities used for responses to equating section fn would be similar to the distributions shown for the first two samples. As can be seen, the mean of the distorted abilities is about 1/3 of a standard deviation below the means of the other two samples, and the percentiles are offset in a manner similar to that found in the real-data estimated ability distributions, although somewhat more exaggerated.

Insert Table 5 about here

Estimation of Item Parameters and Abilities

As before, LOGIST was used to calibrate all items and abilities in a single concurrent execution, with equating items fn used as an anchor to set the scale. The N = 7500 and n = 144 data matrix can be represented as follows:

Items	fn	3ASA3-5	3ASA3-5A	3ASA3-5B
People	n=24	n=60	n=30	n=30
Sample 1	x	x		
Sample 2	x		x	
Sample 3	x			x

The above design produces two different sets of item parameter estimates for the total 60-item test, one as part of 3ASA3-5, and the second as part of the combination of 3ASA3-5A and 3ASA3-5B. For Sample 3, where the true abilities differ for responses to fn items and 3ASA3-5B items, only one ability estimated is produced from the unidimensional IRT model.

Scaling of Estimates

As before, the results of this LOGIST calibration are not on the same scale as the true item and person parameters. The same type of scaling transformation as used in Simulation 1 was repeated here, using the estimated and true abilities for Sample 1.

Comparison of Estimated and True Parameters

Summary statistics for the estimates of item and person parameters after the scaling transformation are presented in Table 5. The percentile comparisons among distributions of estimated abilities are graphically displayed in Figure 1. Figures 11 compares the estimated item parameters and abilities with true item parameters and abilities for test 'form' 3ASA3-5 and Sample 1. Figure 12 compares only the estimated item parameters with true item parameters for the total test 'form' 3ASA3-5A+5B, constructed by combining the items from 3ASA3-5A and 3ASA3-5B. Ability estimates are not compared in Figure 12 since there are two true abilities for Sample 3.

Insert Figures 11 and 12 about here

For the intact form 3ASA3-5, Figure 11 shows the a 's to be slightly underestimated, although the mean estimated a is the same as the mean true a . The b 's are very well estimated. The c 's are about as well estimated as one typically sees, as are the abilities. For the 'pretest' form 3ASA3-5A+5B, shown in Figure 12, the a 's are slightly overestimated. The b 's are greatly overestimated; and the c 's are slightly overestimated.

The explanation for the phenomena exhibited by the 'pretest' form is relatively simple. The individuals in Sample 2 respond to both fn items and pretest items with the same true ability. However, most of the individuals in Sample 3 respond to fn items with one true ability, and to pretest items with a lower true ability. The number of items is roughly the same in both instances (24 for fn and 30 for the pretest). Thus LOGIST will, as much as possible, produce an estimated ability for simulees in Sample 3 that is somewhere in between the two true abilities. This estimate will be higher than the true ability with which responses were generated to the pretest items. A person will get a pretest item incorrect more frequently than is expected on the basis of this ability estimate. Therefore, the estimation procedure behaves as if the pretest item is more difficult than it really is. The unidimensional estimation procedure is given incorrect information from the data as to the item location.

Wingersky and Lord (1984) show that for middle difficulty items, the sampling correlation between estimated a 's and estimated b 's is positive. If the b 's are overestimated, the a 's will also be overestimated. Wingersky and Lord also show that, for middle difficulty items, the sampling correlation between estimated a and estimated c is positive. Thus, if the a 's are overestimated, then so are the c 's on average.

Summary statistics are presented for the estimated abilities in Table 5 and Figure 1. It is interesting to note that the estimation procedure produces estimated abilities for Sample 3 that are not much different from those estimated for Samples 1 and 2. The difference in true ability means disappears. Although there are still differences in each percentile point recorded, these differences are smaller than those modeled with the true abilities. Part of this is due to the production of ability estimates for Sample 3 that lie between the two true abilities, but the extent of differences was a surprise. Because the model does not incorporate two ability dimensions, the differential item responses are reflected mostly in the estimated item difficulties, and not in the estimated abilities. As a result, these estimated abilities do not have a relationship across samples that is very similar to that seen for the estimated abilities for real data shown in Table 4 and Figure 10.

Equating Results

The impact of this type of simulated multidimensionality on equating is seen in the top two plots in Figure 13, where equating and residual plots resulting from the equating of 'pretest' form 3ASA3-5A+5B to operational form 3ASA3-5 are depicted. As expected, this type of lack of model fit has a large impact on equating. At some points on the raw score metric, the differences between scaled scores is over 30 scaled score points. Table 3 shows that there is a difference of about 25 points in the scaled score means, as well as about a 7 point difference in the scaled score standard deviations.

Insert Figure 13 about here

These differences are much larger than the largest differences among scaled score means and standard deviations found with real data. There, the maximum difference between a preequating scaled score mean and a criterion mean was +13 scaled score points. The associated difference between scaled score standard deviations was +3.0 scaled score points. However, in contrast with the earlier simulation study, the mean difference found here is IN THE SAME DIRECTION AS THAT FOUND IN REAL DATA.

The equating differences are explainable in terms of the item parameter miss-estimations previously described. Figure 14 compares the estimates of item parameters from the pretest form against the operational form. Different plotting symbols are used to indicate whether an item comes from pretest 3ASA3-5A or 3ASA3-5B. As expected, parameter estimates from 3ASA3-5B, with the simulated multidimensionality, cause the a 's and c 's to be slightly higher for the pretest form. Table 5 shows that the pretest mean a is 1.01 compared to the final form mean a of .98; the pretest mean c is .15 compared to the final form mean c of .13. The item difficulties are substantially overestimated; the pretest mean b is .26 while the final form mean b is .04.

Insert Figure 14 about here

The resulting impact on equating is most easily seen in Figure 15, which contains plots of the test characteristic curves for the two simulated forms. Because of the overestimation of the item difficulties, the test characteristic curve for the pretest form is shifted to the right of the final form. For true ability levels above -1.0, the number right true score on this form will be lower than on the final form. The pretest form appears more difficult for these examinees. Note, however, that for examinees with very low true ability, the pretest form actually appears easier.

Insert Figure 15 about here

Equating of Estimates to True Values

It is again instructive to examine the equatings of each simulated test form to the true test form. In this way, we can isolate and study estimation errors separately for each form.

The bottom two sets of plots in Figure 13 show equating results when 3ASA3-5 and 3ASA3-5A+5B are equated to the true test form 3ASA3. The resultant mean scaled scores and standard deviations are shown in Table 3. The differences seen for form 3ASA3-5 again indicate the magnitude of equating differences that can be explained on the basis of what is predominantly the imprecision of the estimation procedure alone, since the distribution of true ability for the sample taking this form was the same as the true distribution of ability. It is reassuring to note, through a comparison with plots in Figure 9, that the multidimensionality simulated for items not contained in this 60 item set has a negligible impact on equating errors for this form. The bottom plots in Figure 13, depicting the equating of the simulated pretest form to the true form, demonstrate the impact on equating when the data do not fit the model.

Conclusions from the Second Simulation

The multidimensionality modeled in this simulation was designed to reflect certain intuitively justifiable hypotheses. It is clear that when compared to results with real data, the model is greatly exaggerated. It is also clear, from the resulting distributions of estimated abilities, that while this model may be a step closer than the first simulation to explaining real data results, it is by no means complete.

Third Simulation: Mean Differences and Multidimensionality

The advantage of simulation studies is that they can be used not only to isolate phenomena of interest, but also that they can be used to study controlled combinations of such phenomena. The results of the second simulation study were dissimilar to real-data results in an important way: the relationship among the distributions of estimated abilities did not resemble very closely the relationships found with real data. This third simulation attempts to model the real-data results more faithfully, by combining the phenomena studied in the first two simulations.

Simulated Samples

Using the true abilities, three more samples of N = 2500 each were drawn. The first two samples were drawn with no modification to the true ability distribution. The third sample was drawn after decreasing the mean true ability by .35, as in the smallest mean decrease in the first simulation study.

The 60 items in 3ASA3 were considered to be intac' form 3ASA3-6, and were taken, along with equating section fn, by the first sample. The same random 30-item subset as 3ASA3-5A is considered here to be 3ASA3-6A, and was taken, along with equating section fn, by the second sample. The remaining random subset of 30 items, 3ASA3-5B, is considered here to be 3ASA3-6B, and was taken, along with equatirg section fn, by the third sample. When the third sample responds to the 24 items in the equating section, it does so with average true ability decreased by .35. When the third sample responds to the 30-item 3ASA3-6B, the average true ability is decreased by .35 and then THE SAME distortion in true abilities as described earlier is repeated. Note that this distortion is applied to the distribution of true abilities AFTER the mean true ability has been decreased.

The results of the generation of samples of true abil'.y are shown in the bottom portion of Table 6. For the third sample, only the distorted abilities used to generate responses to 3ASA3-6B are shown. The mean-shifted true abilities used to generate responses to equating section fn would be similar to that of the other two samples except for the shift, and also to sample 2 from the first simulation study (see Table 2). The mean of the distorted true abilities is now about 2/3 of a standard deviation below the other true sample means, as opposed to the 1/3 obtained from distortion alone (see Table 5).

Insert Table 6 about here

Estimation of Item Parameters and Abilities

The LOGIST calibration of items and abilities is a single concurrent run, as in the previous simulations. The design of this N = 7500 and n = 144 calibration is the same as that for the second simulation. For completeness, it is repeated here.

Items	fn	3ASA3-6	3ASA3-6A	3ASA3-6B
People	n=24	n=60	n=30	n=30
Sample 1	x	x		
Sample 2	x		x	
Sample 3	x			x

As before, two diff.rent sets of 60-item total test parameter estimates are obtained, one as part of 3ASA3-6, and the second as part of the combination of 3ASA3-6A and 3ASA3-6B. Only one estimate of ability is obtained fcr individuals in Sample 3.

Scaling of Estimates

As before, the results of this calibration must be transformed to the scale of the true item parameters before any comprarisons can be made. The same type of transformation as before was performed, using the true and estimated abilities for Sample 1.

Comparison of Estimates and True Parameters

Summary statistics for the estimation of item and person parameters after the scaling transformation are presented in Table 6 and Figure 1. As before, the distortion of true abilities has caused the average estimated b on the 'pretest' to be larger than that of the intact form, .18 vs. 0.0. However, the difference in the averages is less than that seen in Table 5 for distortion alone, where the means were .26 and .04. The mean estimated a 's are identical as are the mean estimated c 's.

Figures 16 and 17 graphically compare the estimated and true item parameters for the two 60-item forms, 3ASA3-6 and 3ASA3-6A+6B. A comparison of these figures with Figures 11 and 12 shows that the a 's for the both forms are better estimated here, as are the c 's. As expected, the estimated item difficulties for the intact form appear as well estimated as before; those for the pretest form are less overestimated.

Insert Figures 16 and 17 about here

The summary statistics f.r the estimated abilities in Table 6 and Figure 1 show that the mean estimated ability for sample 3 is about 1/3 of a standard deviation below that of the other two samples, in contrast to the near equality seen in Table 5. In addition, the percentiles reflect the distortion in true abilities to a much greater degree. The differences among the distributions of estimated ability in Table 6, with mean shift and multidimensionality introduced, in contrast to Table 5, with multidimensionality alone, provide some intuition as to the behavior of the item parameter estimates.

The third sample in each simulation takes two blocks of items, equating section fn in common with the other two samples, and the second block of pretest items. The only information for estimating the item parameters for this second set of pretest items comes from the third sample in each case. With multidimensionality introduced alone, the responses to items in equating section fn by the third sample are equivalent to those from the other two samples, since all three are samples from the same distribution of true ability. Therefore, the multidimensionality introduced into the responses for the second block of pretest items is reflected in the item parameter estimates for those items alone, and is not attributed to differences in ability. In contrast, when a mean shift in ability and multidimensionality are introduced, the third sample responds to equating section fn with a lower mean ability. Therefore the lack of success on the second block of pretest items introduced by the multidimensionality can be attributed in part to the lower mean true ability. Therefore, the item difficulties are less overestimated.

Although somewhat more exaggerated, the relationships between item parameter estimates for the pretest and intact forms shown in

Table 6 and Figure 1 replicate those found with real data. In addition, the relationships among the distributions of estimated abilities shown in Table 6 and Figure 1 are similar to those found in Table 4 and Figure 10 for the four real-data pretest samples contributing the largest number of items to final form 3ASA3.

Equating Results

The impact on equating of a mean shift in ability and the introduction of multidimensionality is shown in equating and residual plots at the top of Figure 18. These plots depict the equating of pretest form 3ASA3-6A+6B to the intact form 3ASA3-3. As expected, the impact on equating is large, although not as large as that for multidimensionality alone. Table 3 shows that the impact has been reduced to about 22 scaled score points at the mean, in contrast to 25 scaled score points for multidimensionality alone.

Insert Figure 18 about here

As before, the equating differences are explainable in terms of item parameter miss-estimations previously described. Figure 19 graphically compares the two sets of estimates, with different plotting symbols indicating the membership of an item in 3ASA3-6A or 3ASA3-6B. The previous remarks made in reference to the plots in Figure 14 are applicable here.

Insert Figure 19 about here

The test characteristic curves for the two forms are shown in Figure 20. While similar to Figure 15, the curves are closer to each other, particularly for low ability levels.

Insert Figure 20 about here

Equating of Estimates to True Values

The bottom two sets of plots in Figure 18 show equating results when 3ASA3-6 and 3ASA3-6A+6B are equated to the true test form 3ASA3. The resultant mean scaled scores and standard deviations are shown in Table 3. The differences seen for the intact form again reflect the magnitude of equating differences that can be explained on the basis of what is predominantly the imprecision of the estimation procedure alone. It is again reassuring that this equating is not contaminated by the introduction of a mean shift and multidimensionality in the third sample. The bottom set of plots in Figure 18 demonstrates the impact on equating when both phenomena are introduced.

Conclusions from the Third Simulation

The introduction of a slight decrease in the mean of the true abilities in conjunction with a certain type of multidimensionality produces results that are consistent with those seen in real data. However, it is clear that the effects are exaggerated when compared to the real data results. Presumably this is because the multidimensionality was modeled for every individual with a particular true ability in exactly the same way. A more realistic model would introduce this type of multidimensionality for only a certain proportion of individuals with the same true ability. It is likely that this modification would produce results that resemble real-data results even more closely.

CONCLUSIONS

The purpose of this study was to understand the impact of differences in true ability in a particular application that depends upon item parameter invariance: preequating. Starting from reasonable hypotheses suggested by the real SAT-mathematical preequating data, three controlled simulations were conducted to test these hypotheses. The results clearly have implications beyond an understanding of a particular set of real data. They can be stated generally as follows:

1. Differences in mean true ability can cause differences in the precision with which a particular estimation procedure estimates parameters, even when the data fit the particular IRT model used. The effect of this differential precision on preequating a test is relatively moderate. The particular differences in ability studied here produced the opposite effect on preequating than what was expected, based on the real data preequating, although this could have been predicted in advance.

2. The introduction of a particular kind of multidimensionality in the data can have a large impact on estimation precision when the IRT model is unidimensional. The computer program used here, LOGIST, reflects the impact of this type of multidimensionality mostly in the item parameter estimates, rather than the ability estimates.

3. The combination of a slight decrease in mean ability and a particular type of multidimensionality in the data also has a large impact on estimation precision when the IRT model is unidimensional, although the impact is lessened somewhat. This occurs because the lack of model fit is incorporated into the estimated abilities as well as the item parameter estimates.

In keeping with the desire to understand the particular set of SAT-mathematical data that generated the need for these simulation studies, one conclusion can be stated more specifically:

> Based on the reasonable simulations studied here, poor preequatings obtained for the particular set of SAT mathematical data were consistent with a combination of a slight decrease in mean true ability, and a particular type of multidimensionality introduced into specific pretest sections. Regardless of the causes to which one wants to attribute this multidimensionality, this conclusion appears

inescapable. Given sufficient time and money, it is likely that further simulations could be devised that are even more consistent with real-data results than those presented here.

What are the implications of these conclusions for future efforts to capitalize on the invariance properties of true item and person parameters? There are at least three:

1. The unidimensional IRT model parameter estimates produced by LOGIST are relatively immune to imprecisions due to small differences in true ability. Differences as large as a standard deviation begin to have a greater impact, and the importance of that impact will clearly depend upon the particular application for which invariance is desired. Vertical equating applications, where differences may be as large or larger than those studied here, should be approached with caution.
2. If data do not fit the unidimensional model in the particular manner modeled here, LOGIST provides some indication of this through the production of inconsistent results, e.g., the 'failure' of the preequating with real data.
3. Greater efforts must be made both to ensure the data fitted with a unidimensional model are in fact unidimensional, and to develop practical, useful, and informative multidimensional models for the future.

REFERENCES

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT-verbal and mathematical sections* (Research Report 85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating* (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Kingston, N. L., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement*, *9*, 281-288.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service.

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

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 136-156.

Wingersky, M. S. (1985). Personal communication.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error on certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST V users guide. Princeton, NJ: Educational Testing Service.

Table 1

Summary Statistics for True Item and Person Parameters,

Test Form 3ASA3 and Equating Section fn

True Item Parameters

	3ASA3	fn
Max a	1.71	1.33
Mean a	.98	.94
Median a	.92	.94
Min a	.30	.43
S.D. (a)	.33	.25
n	60	24
Max b	2.33	2.44
Mean b	-.01	.17
Median b	.05	.25
Min b	-3.32	-3.25
S.D. (b)	1.27	1.27
n	60	24
Max c	.41	.29
Mean c	.14	.14
Median c	.13	.12
Min c	0	0
S.D. (c)	.10	.08
n	60	24

True Abilities

					Percentiles				
N	Mean	SD	Min	Max	10	25	50	75	90
2744	-.01	1.03	-7.35	3.91	-1.36	-.66	.01	.71	1.24

Table 2

Summary Statistics for First Simulation Study: Mean Shift Only

All Parameter Estimates Have Been Transformed to the Scale of the True Values

Item Parameter Estimates

Test Form:	3ASA3-1	3ASA3-2	3ASA3-3	3ASA3-4	fn
Sample:	1	2	3	4	all samples
max a	1.72	1.72	1.72	1.72	1.41
mean a	.95	1.01	.99	1.05	.96
median a	.91	.97	.92	1.01	.96
min a	.33	.30	.37	.36	.47
S.D. (a)	.32	.33	.34	.33	.27
n	60	60	60	60	24
max b	2.67	3.02	2.91	2.26	2.62
mean b	.01	.05	.03	-.01	.19
median b	-.04	.07	-.04	.03	.25
min b	-2.96	-2.95	-3.38	-3.53	-3.04
S.D. (b)	1.28	1.24	1.28	1.21	1.24
n	60	60	60	60	24
max c	.40	.40	.37	.36	.28
mean c	.12	.13	.12	.13	.12
median c	.09	.12	.11	.12	.11
min c	0	0	0	0	.01
S.D. (c)	.10	.10	.09	.09	.08
n	60	60	60	60	24

Ability Estimates

	Form										
Sample	Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-1(60)	2498	-0.01	1.04	-7.92	4.27	-1.27	-0.65	-0.02	0.69	1.31
2	3ASA3-2(60)	2498	-0.31	1.01	-7.92	4.10	-1.53	-0.92	-0.28	0.36	0.90
3	3ASA3-3(60)	2500	-0.67	1.05	-7.92	3.49	-1.83	-1.25	-0.67	-0.02	0.59
4	3ASA3-4(60)	2500	-0.98	0.99	-7.92	2.60	-2.15	-1.58	-0.96	-0.34	0.20

True Abilities

	Form										
Sample	Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-1(60)	2500	-0.02	1.02	-7.35	3.85	-1.34	-0.67	-0.03	0.71	1.26
2	3ASA3-2(60)	2500	-0.35	1.00	-3.78	3.56	-1.67	-0.97	-0.32	0.35	0.86
3	3ASA3-3(60)	2500	-0.72	1.02	-8.05	3.15	-2.04	-1.37	-0.73	0.01	0.56
4	3ASA3-4(60)	2500	-1.07	1.01	-5.20	2.80	-2.36	-1.75	-1.08	-0.35	0.18

Table 3

Scaled Score Means and Standard Deviations Resulting

from Equatings with Simulated Data

Equating Pairs	Scaled Score Mean	Scaled Score Standard Deviation	Scaled Score Mean Minus True Scaled Score Mean
3ASA3 (true)	485	113	
3ASA3-2→3ASA3-1	486	112	1
3ASA3-3→3ASA3-1	486	112	1
3ASA3-4→3ASA3-1	480	108	-5
3ASA3-1→3ASA3 (true)	487	111	2*
3ASA3-2→3ASA3 (true)	488	110	3
3ASA3-3→3ASA3 (true)	488	110	3
3ASA3-4→3ASA3 (true)	482	106	-3
3ASA3-5A+5B→3ASA3-5	510	120	25
3ASA3-5→3ASA3 (true)	487	112	2*
3ASA3-5A+5B→3ASA3 (true)	512	119	27
3ASA3-6A+6B→3ASA3-6	507	120	22
3ASA3-6→3ASA3 (true)	487	111	2*
3ASA-6A+6B→3ASA3 (true)	509	119	24

*Difference is due predominantly to errors of estimation.

Table 4

Summary Statistics for Estimated Abilities from Real Preequating Data for Test Form 3ASA3, Sorted by Median Estimated Ability

Estimated Abilities*										
						Percentiles				
Sample (n)**	N	Mean	S.D.	Min	Max	10	25	50	75	90
X316(2)	2704	.19	1.01	-7.33	3.56	-.95	-.35	.24	.82	1.32
X313(1)	2795	.17	1.03	-7.33	3.49	-.91	-.38	.23	.76	1.30
3ASA3-Oper.	2772	.16	.97	-4.19	3.18	-1.14	-.43	.22	.82	1.33
X233(3)	2490	.14	1.04	-7.33	3.89	-1.21	-.54	.19	.84	1.43
X226(1)	2561	.13	1.03	-7.33	3.86	-1.15	-.54	.17	.82	1.39
X232(2)	2522	.15	1.04	-7.33	3.91	-1.13	-.51	.17	.83	1.45
X241(2)	2493	.14	1.00	-4.73	3.50	-1.11	-.51	.16	.84	1.36
X243(4)	2489	.14	1.03	-7.33	3.64	-1.13	-.49	.16	.78	1.39
X405(1)	2514	.06	1.21	-7.33	3.20	-1.31	-.60	.14	.82	1.41
X234(3)	2458	.10	1.04	-7.33	3.65	-1.15	-.54	.12	.76	1.41
X415(1)	2828	-.14	1.25	-7.33	3.25	-1.57	-.77	-.02	.66	1.25
Z515(1)	2513	-.17	1.34	-7.33	3.45	-1.55	-.79	-.07	.59	1.26
C2318(9)	2727	-.06	.98	-7.33	3.05	-1.22	-.67	-.10	.55	1.21
C2314(10)	2619	-.10	.98	-3.99	3.18	-1.35	-.76	-.11	.55	1.13
Z512(3)	2616	-.22	1.23	-7.33	3.51	-1.54	-.84	-.17	.51	1.15
C1613(10)	2963	-.26	1.00	-7.33	4.19	-1.44	-.88	-.26	.41	.95
C1614(7)	2883	-.27	1.02	-7.33	3.09	-1.48	-.92	-.27	.39	.99

*These ability estimates are on a different scale than those contained in all other tables.

**The numbers in parentheses are the number of items contributed by this sample to the total of 60 items in test form 3ASA3.

Table 5

Summary Statistics for Second Simulation Study: Distortion Only

All Parameter Estimates Have Been Transformed to the Scale of the True Values

Item Parameter Estimates

Test Form:	3ASA3-5	3ASA3-5A+5B	3ASA3-5A	3ASA3-5B	fn
Sample:	1	2 and 3	2	3	all samples
max a	1.75	1.74	1.74	1.74	1.38
mean a	.98	1.01	1.05	.97	.94
median a	.94	.90	.95	.80	.92
min a	.35	.32	.37	.32	.46
S.D. (a)	.32	.37	.38	.35	.26
n	60	60	30	30	24
max b	2.64	3.07	2.75	3.07	2.64
mean b	.04	.26	-.07	.59	.19
median b	-.02	.35	.14	.63	.24
min b	-3.36	-3.49	-3.49	-2.80	-3.15
S.D. (b)	1.27	1.38	1.39	1.32	1.28
n	60	60	30	30	24
max c	.47	.38	.38	.38	.27
mean c	.13	.15	.14	.15	.12
median c	.12	.14	.13	.14	.10
min c	0	0	0	0	0
S.D. (c)	.11	.09	.09	.10	.08
n	60	60	30	30	24

Ability Estimates

Sample	Form Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-5(60)	2496	0.03	1.05	-7.29	4.11	-1.25	-0.65	0.01	0.71	1.35
2	3ASA3-5A(30)	2496	0.03	1.07	-7.29	3.80	-1.29	-0.66	0.03	0.70	1.34
3	3ASA3-5B(30)	2499	0.04	1.10	-7.29	4.41	-1.19	-0.59	0.02	0.66	1.30

True Abilities

Sample	Form Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-5(60)	2500	0.01	1.02	-4.15	3.85	-1.32	-0.65	0.01	0.71	1.30
2	3ASA3-5A(30)	2500	0.01	1.02	-4.15	3.85	-1.32	-0.65	0.01	0.71	1.30
3	3ASA3-5A(30)	2500	-0.33	0.92	-4.15	3.85	-1.32	-0.88	-0.39	0.11	0.70

Table 6

Summary Statistics for Third Simulation Study: Mean Shift and Distortion

All Parameter Estimates Have Been Transformed to the Scale of the True Values

Item Parameter Estimates

Test Form:	3ASA3-6	3ASA3-6A+6B	3ASA3-6A	3ASA3-6B	fn
Sample:	1	2 and 3	2	3	all samples
max a	1.73	1.78	1.78	1.78	1.37
mean a	1.00	1.00	1.02	.98	.95
median a	.96	.93	.95	.84	.94
min a	.30	.25	.35	.25	.42
S.D. (a)	.34	.40	.38	.42	.25
n	60	60	30	30	24
max b	2.47	2.63	2.38	2.63	2.40
mean b	0	.18	-.13	.49	.17
median b	-.08	.29	-.20	.70	.22
min b	-3.37	-3.84	-3.84	-2.70	-3.38
S.D. (b)	1.28	1.37	1.40	1.27	1.30
n	60	60	30	30	24
max c	.46	.44	.42	.44	.27
mean c	.13	.13	.13	.13	.12
median c	.13	.11	.11	.12	.10
min c	0	0	0	0	0
S.D. (c)	.10	.10	.09	.10	.07
n	60	60	30	30	24

Ability Estimates

Sample	Form Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-6(60)	2499	0.04	1.03	-7.18	4.05	-1.21	-0.61	0.01	0.70	1.34
2	3ASA3-6A(30)	2499	0.02	1.11	-7.18	3.85	-1.26	-0.66	-0.01	0.70	1.37
3	3ASA3-6B(30)	2498	-0.28	1.07	-7.18	3.83	-1.46	-0.81	-0.29	0.34	0.89

True Abilities

Sample	Form Taken (n)	N	Mean	SD	Min	Max	10	25	50	75	90
1	3ASA3-6(60)	2500	0.02	1.02	-7.35	3.91	-1.29	-0.62	0.01	0.74	1.25
2	3ASA3-6A(30)	2500	0.02	1.02	-7.35	3.91	-1.29	-0.62	0.01	0.74	1.25
3	3ASA3-6A(30)	2500	-0.62	0.88	-7.70	3.56	-1.64	-1.07	-0.72	-0.07	0.30

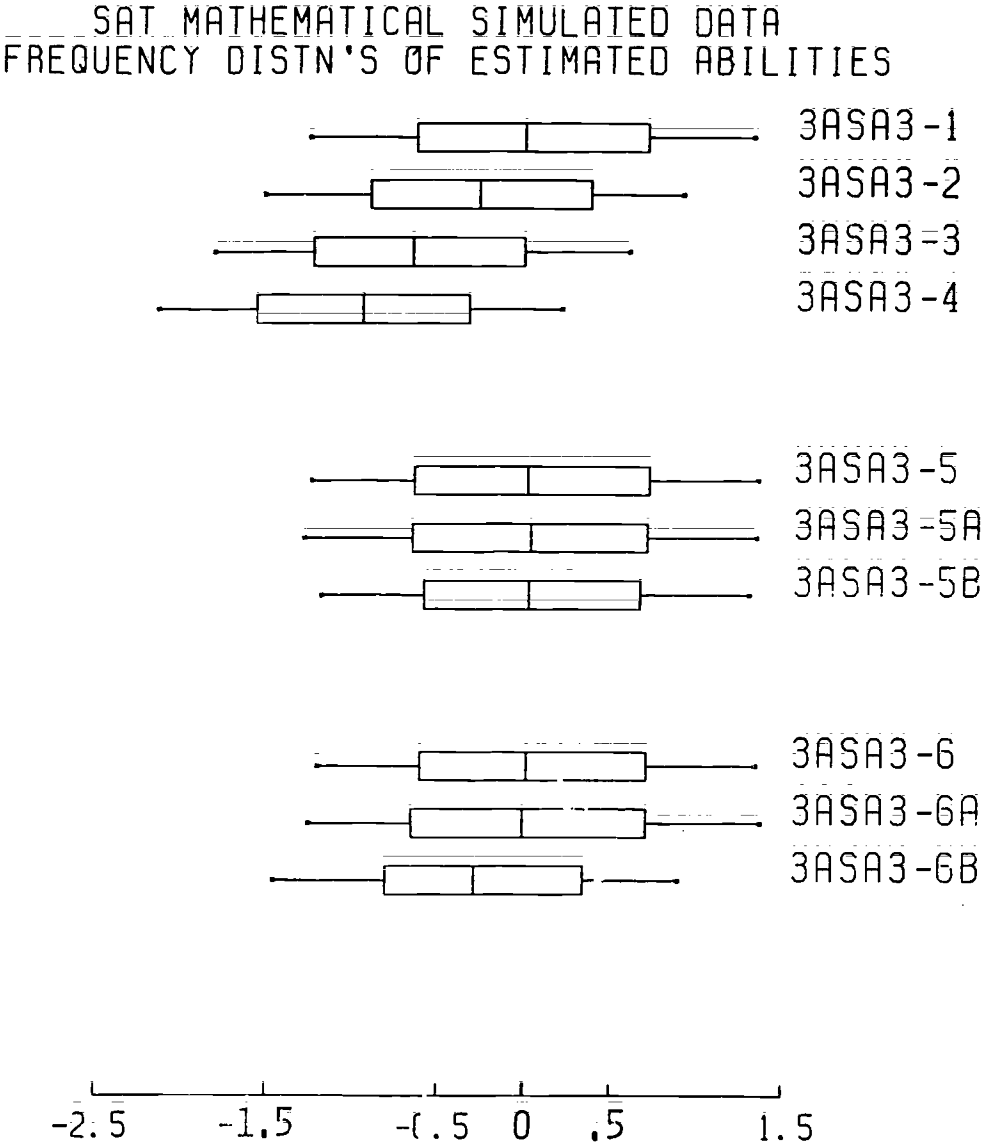

Figure 1: Schematics of summary statistics of distributions of estimated abilities for all simulated samples.

First simulation study: 3ASA3-1, 3ASA3-2, 3ASA3-3, 3ASA3-4.
Second simulation study: 3ASA3-5, 3ASA3-5A, 3ASA3-5B.
Third simulation study: 3ASA3-6, 3ASA3-6A, 3ASA3-6B.

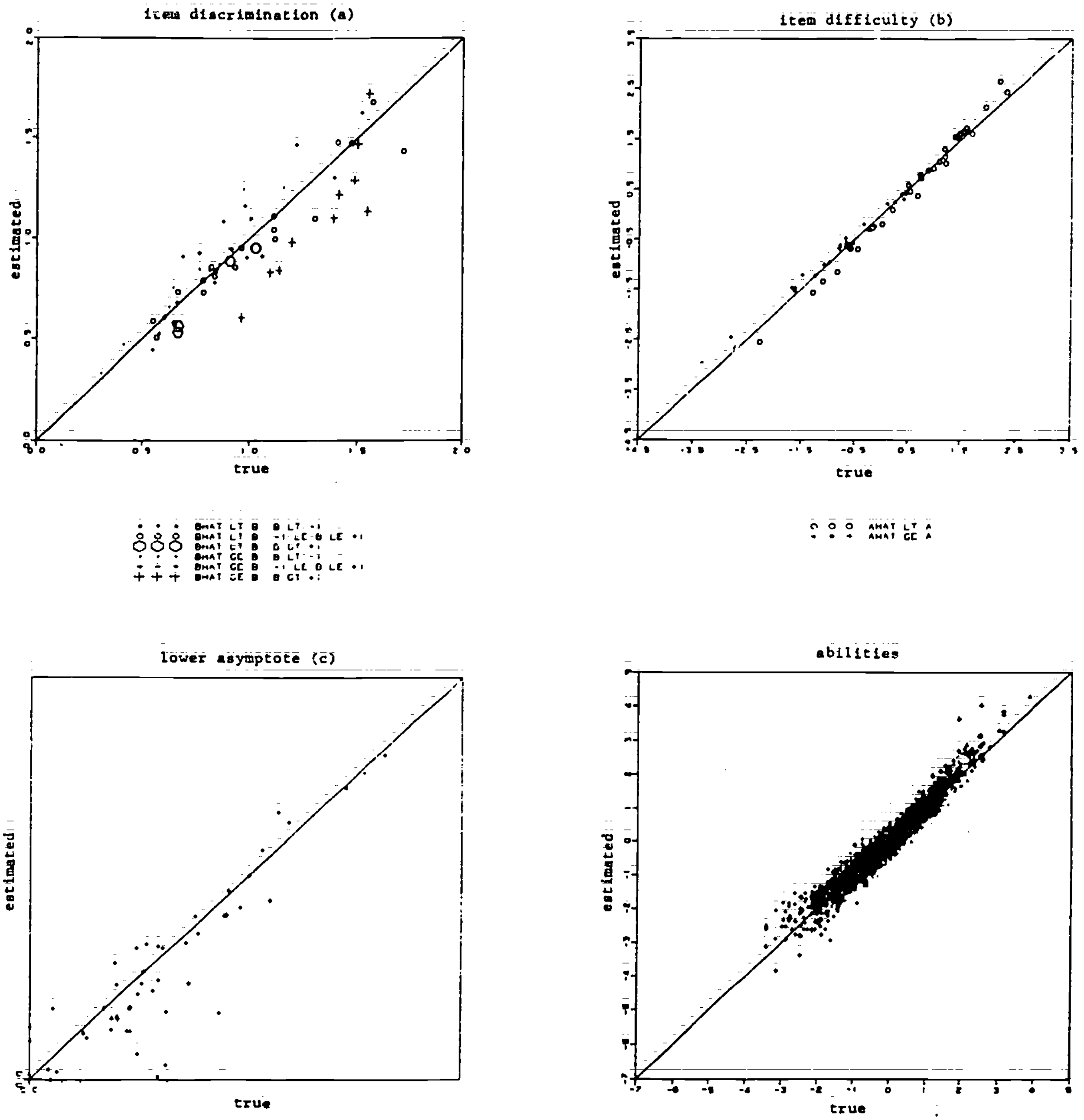

Figure 2: First simulation study: Parameter estimates for Sample 1 taking form 3ASA3-1 vs. true parameter values.

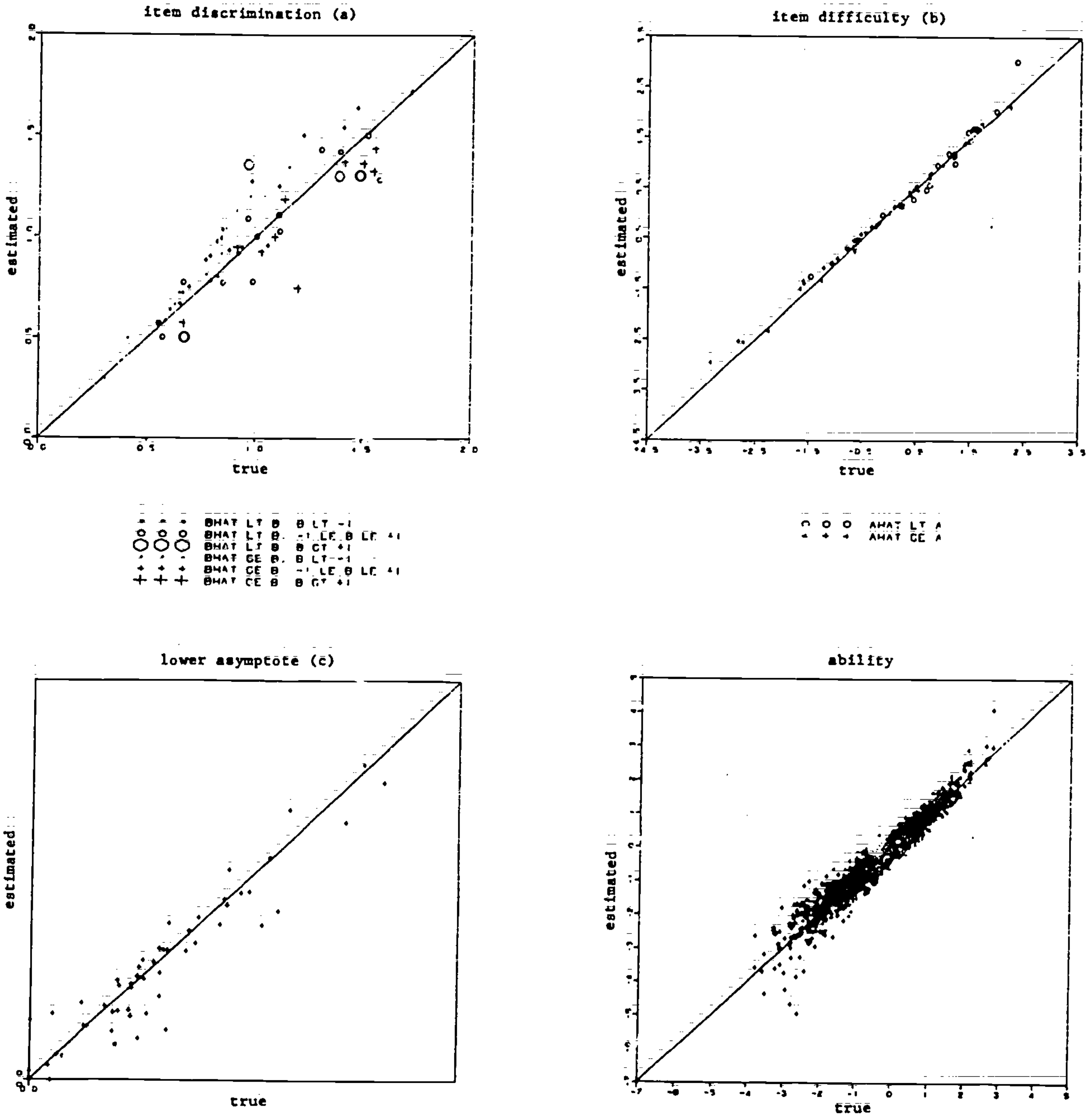

Figure 3: First simulation study: Parameter estimates for Sample 2 taking form 3ASA3-2 vs. true parameter values.

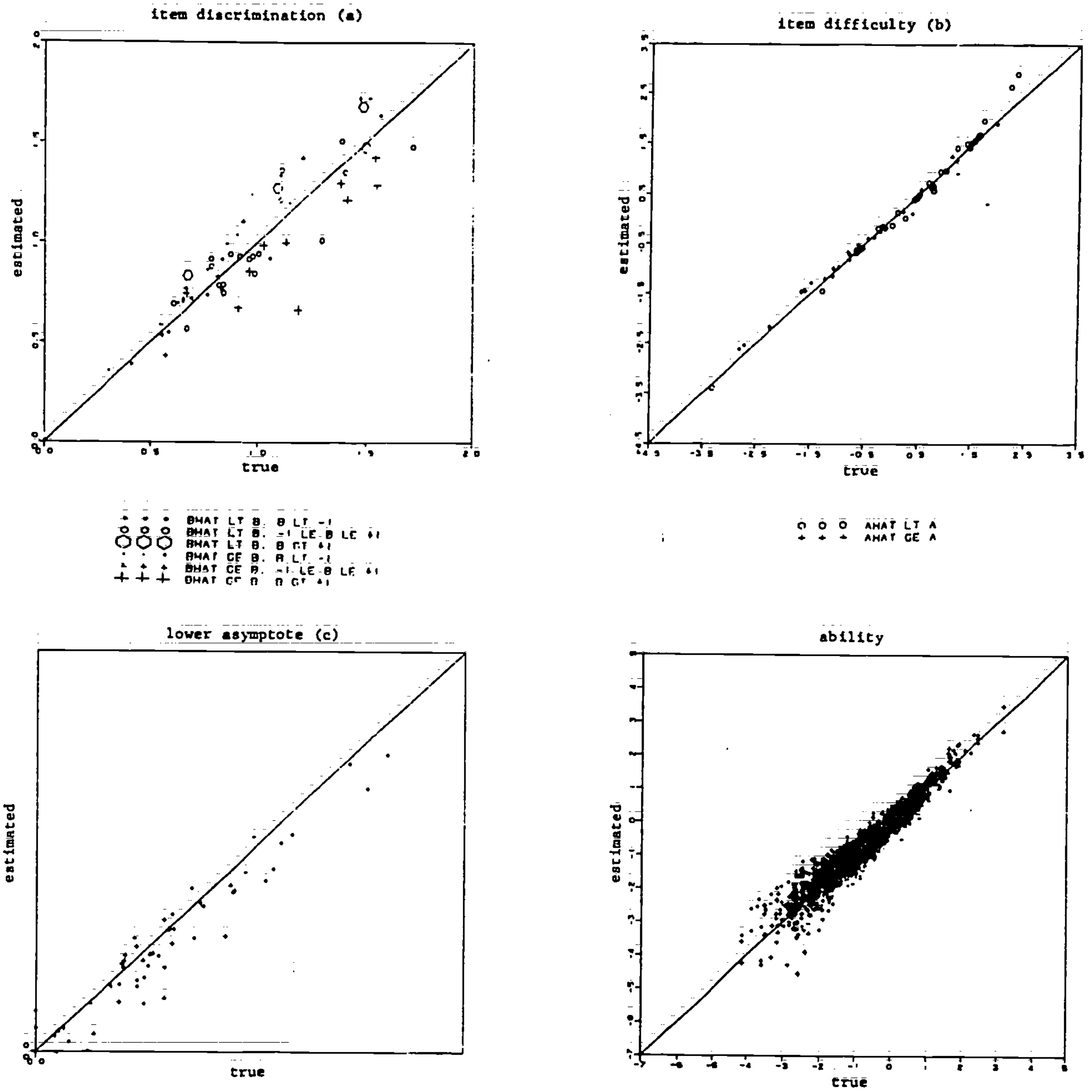

Figure 4: First simulation study: Parameter estimates for Sample 3 taking form 3ASA3-3 vs. true parameter values.

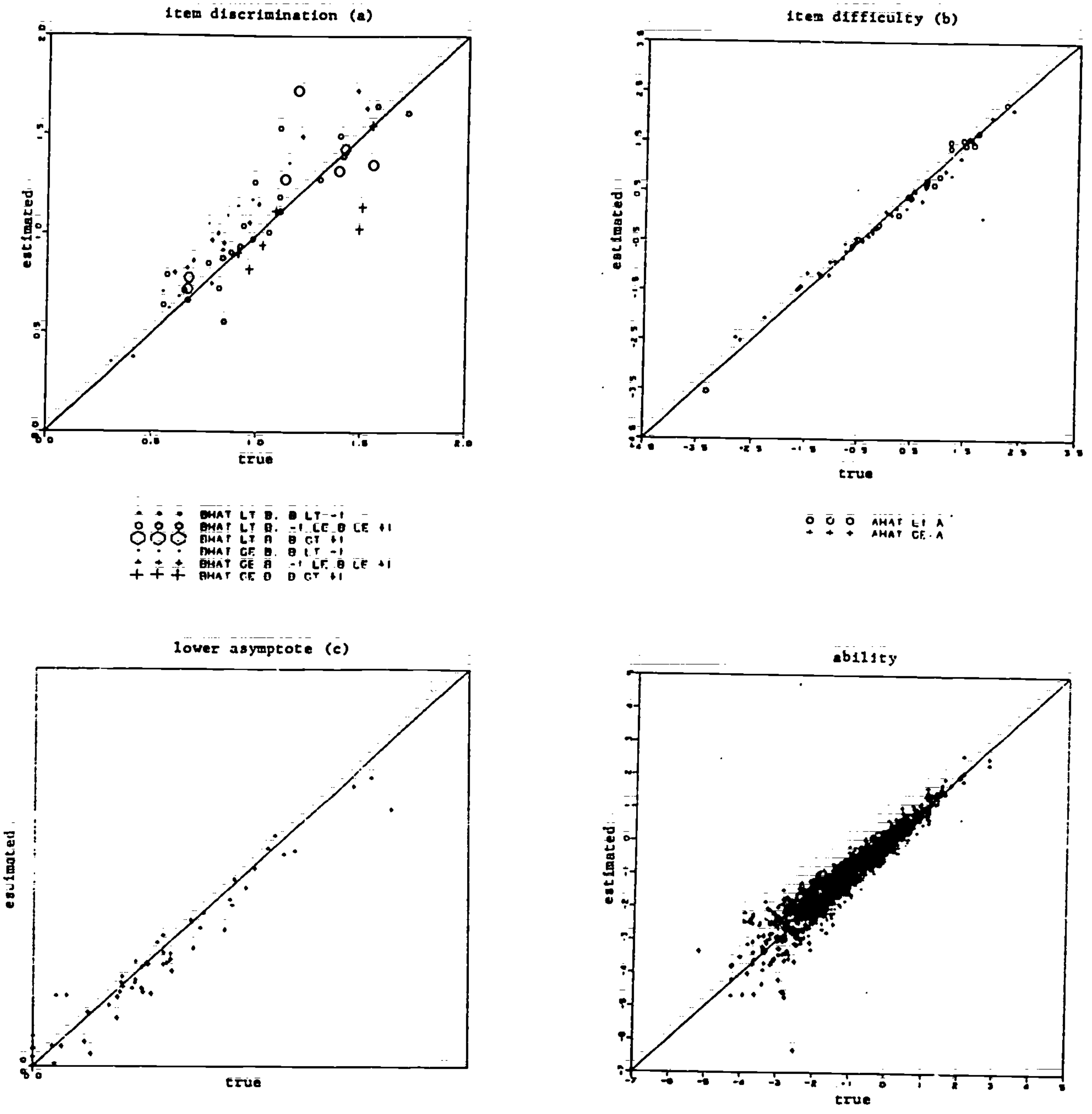

Figure 5: First simulation study: Parameter estimates for Sample 4 taking form 3ASA3-4 vs. true parameter values.

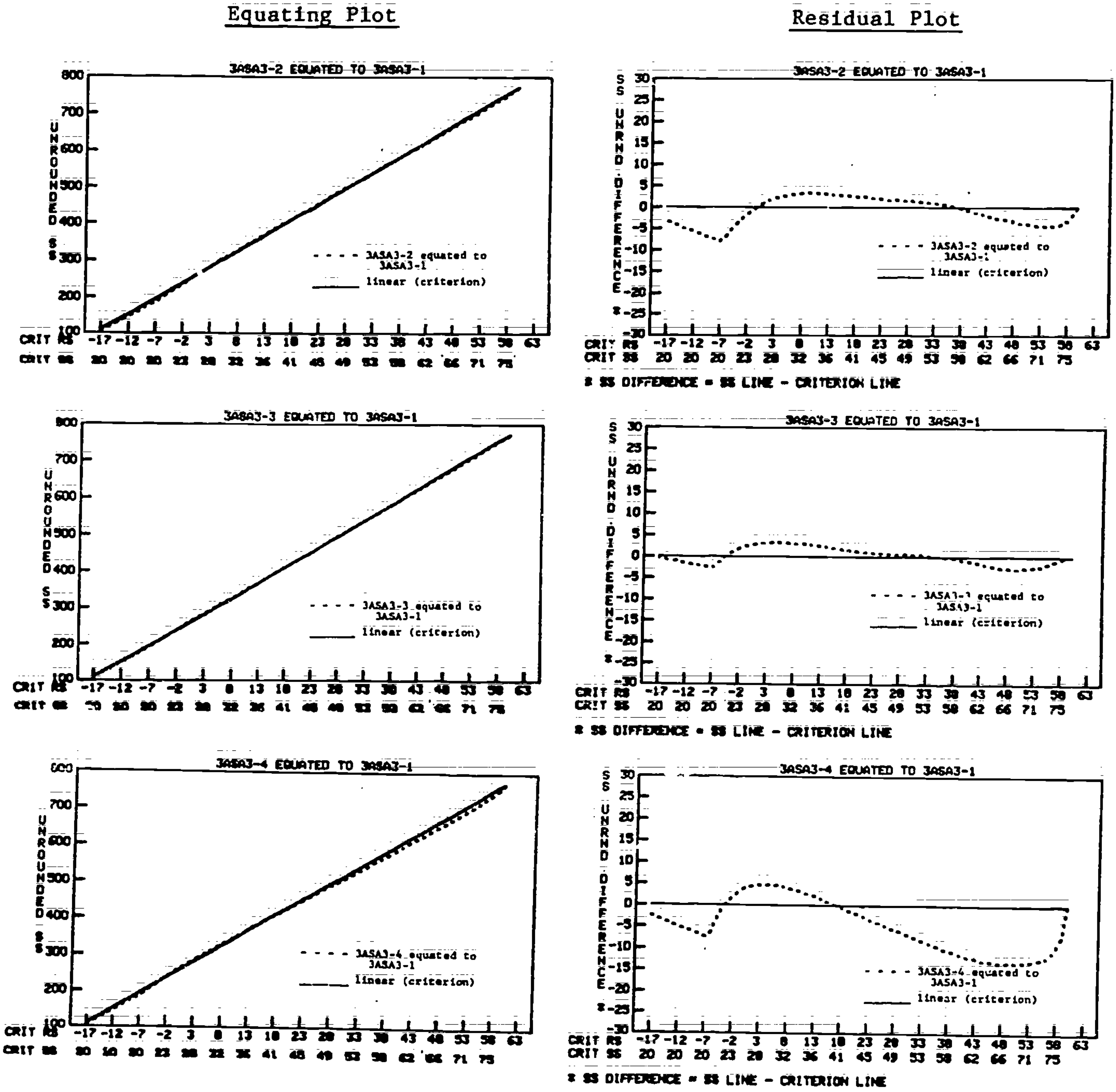

Figure 6. Equating and equating residuals for the first simulation study. Forms 3ASA3-2, 3ASA3-3, and 3ASA3-4 are equated to Form 3ASA3-1.

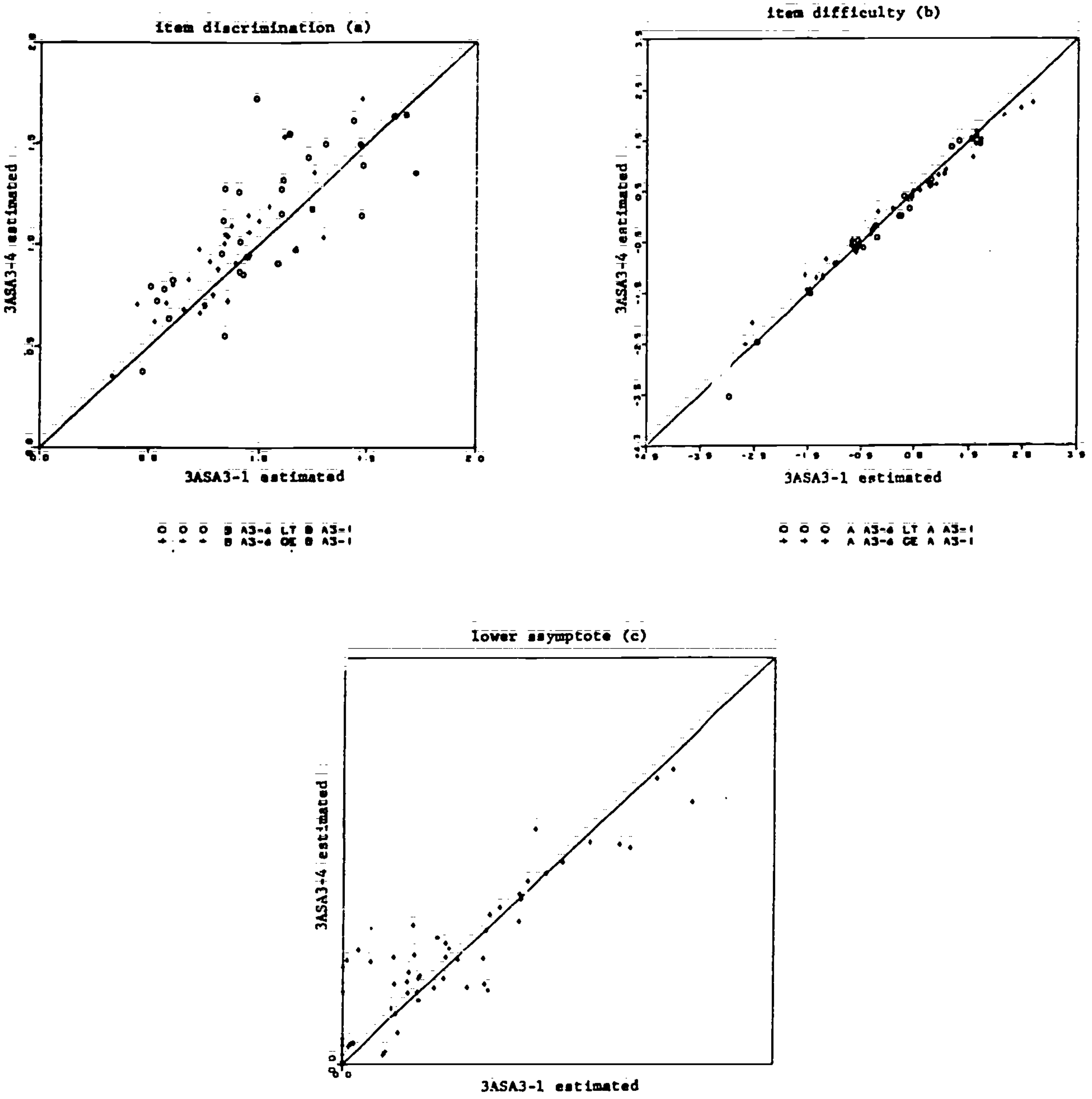

Figure 7. First simulation study: A comparison of 3ASA3-4 estimated parameters vs. 3ASA3-1 estimated parameters.

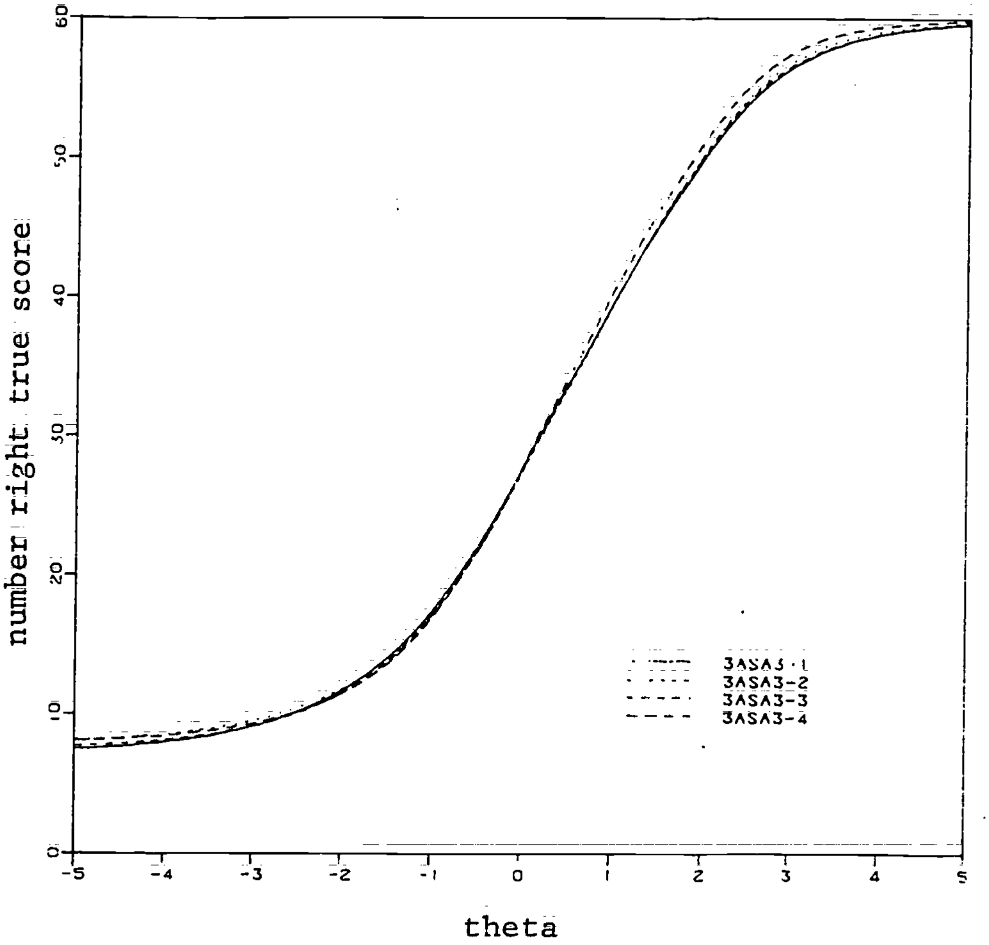

Figure 8. Test characteristic curves for 3ASA3-1, 3ASA3-2, 3ASA3-3, and 3ASA3-4 for the first simulation study.

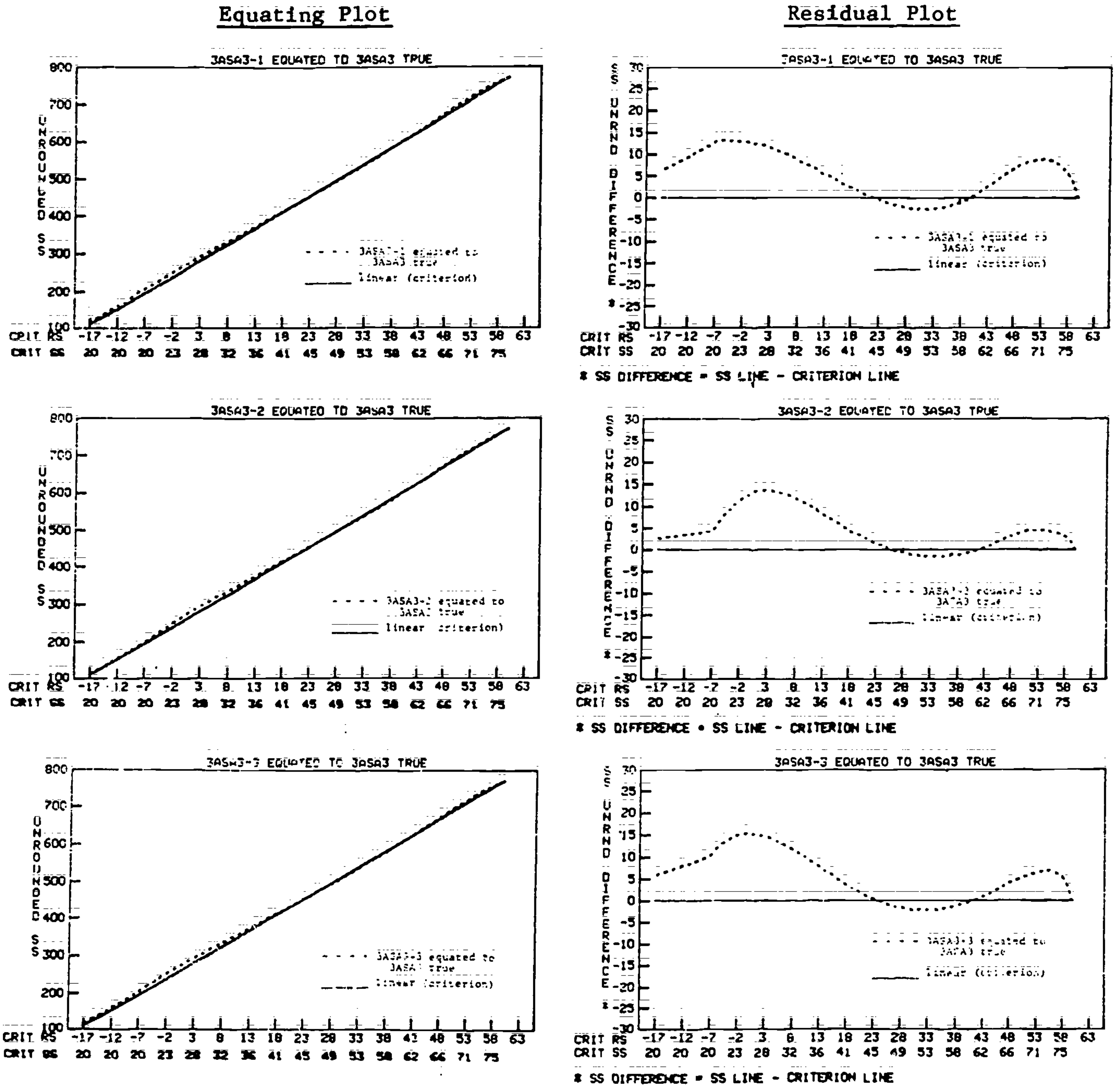

Figure 9. Equating and equating residuals for first simulation study. Forms 3ASA-1, 3ASA-2, 3ASA3-3, and 3ASA3-4 are equated to true form 3ASA3.

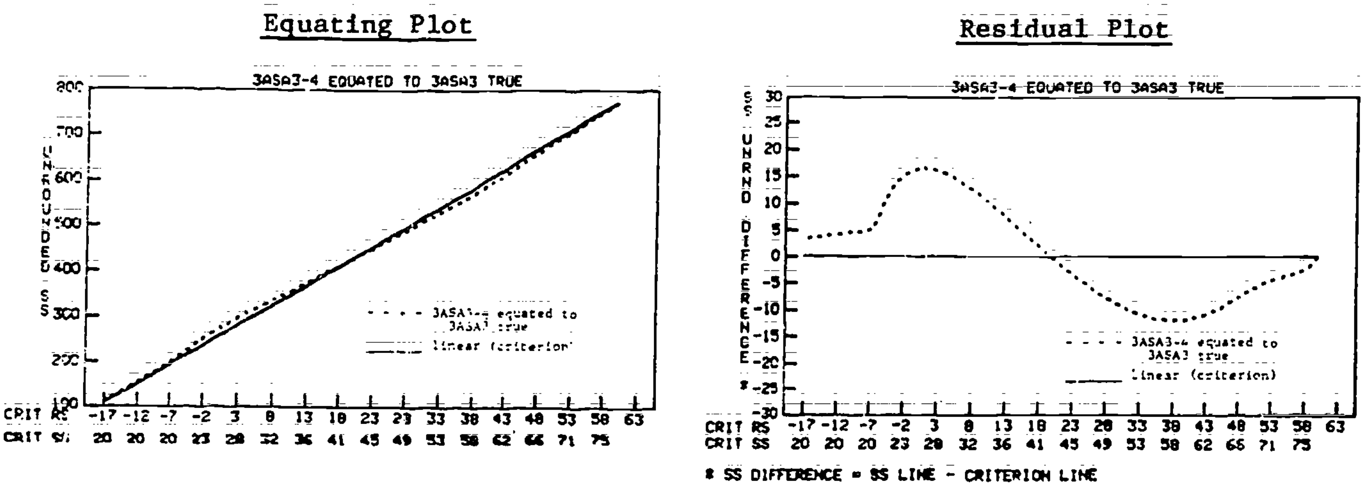

Figure 9 (continued).

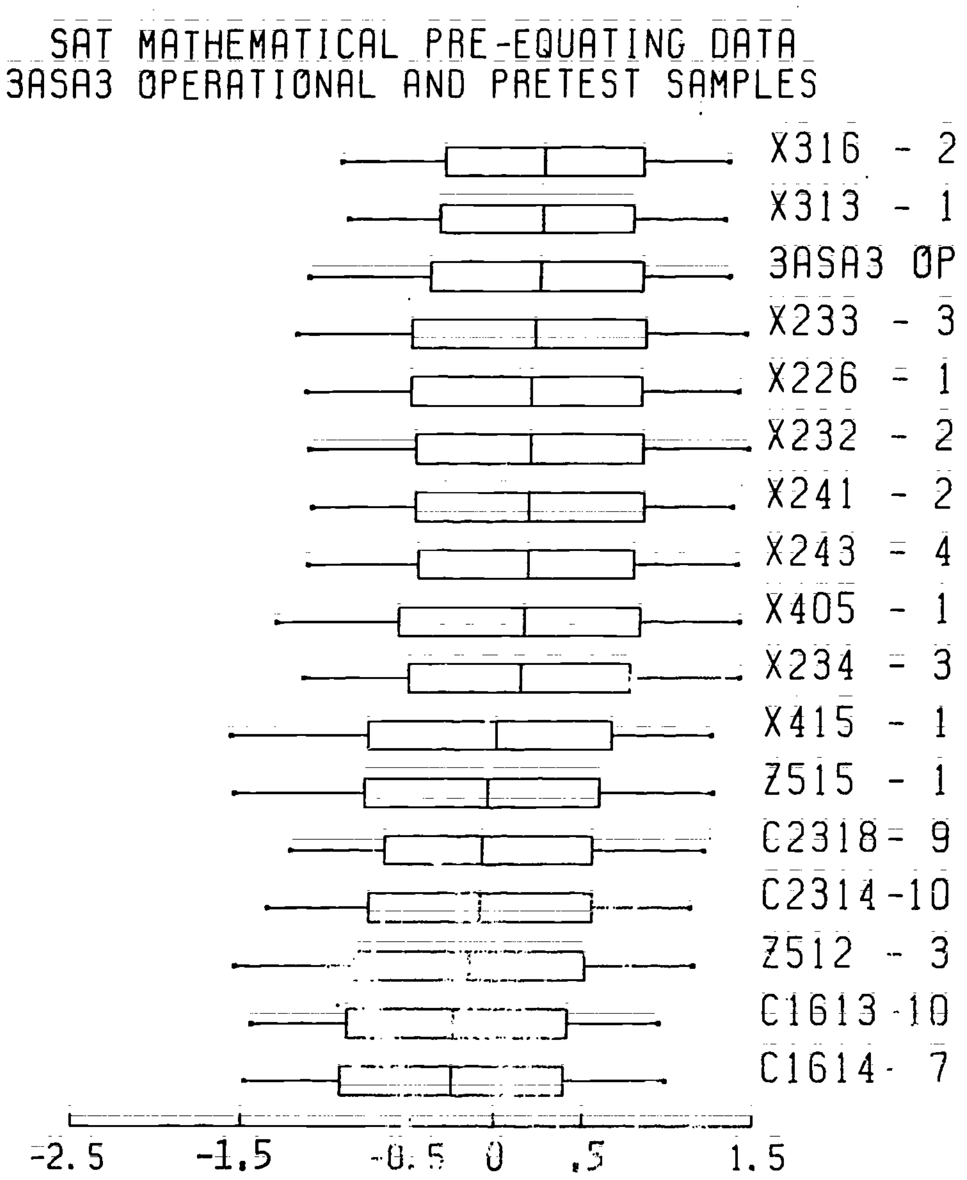

Figure 10: Schematics of summary statistics of distributions of estimated abilities for all pretest samples and one operational sample for SAT mathematical form 3ASA3.

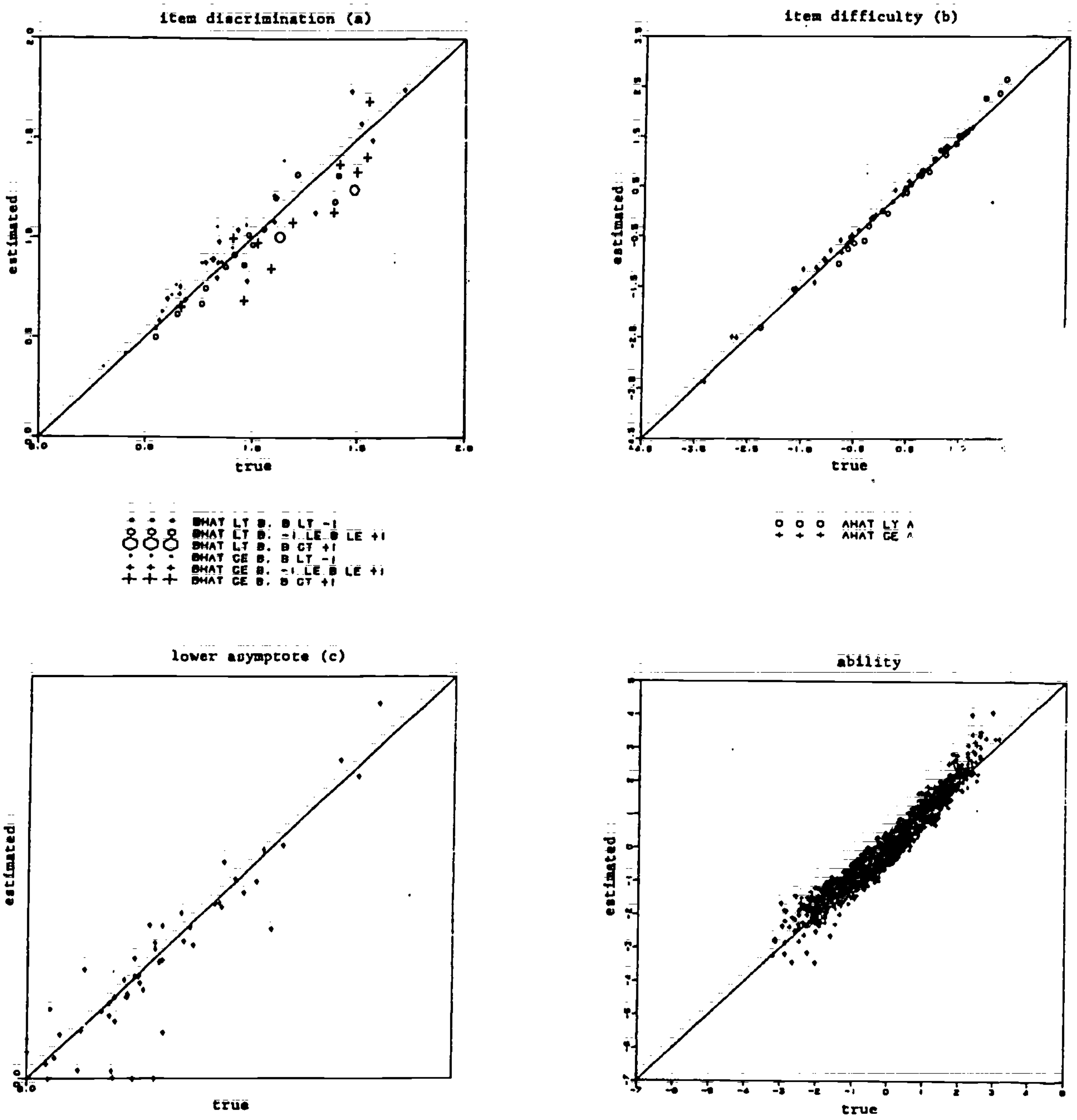

Figure 11: Second simulation study: Parameter estimates for Sample 1 taking form 3ASA3-5 vs. true parameter values.

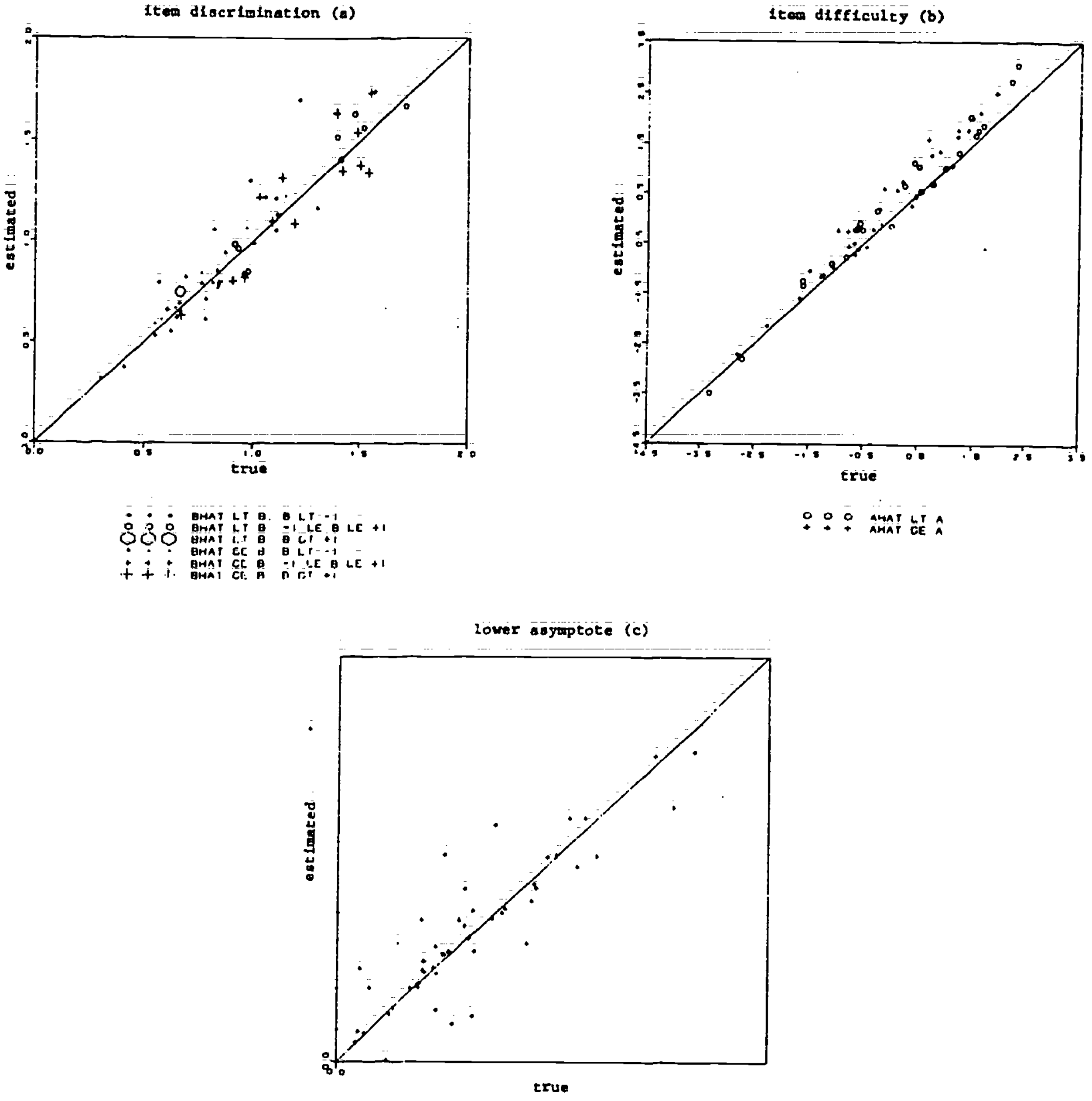

Figure 12. Second simulation study: Parameter estimates for Sample 2 and 3 taking combined form 3ASA3-5A+5B vs. true parameter values.

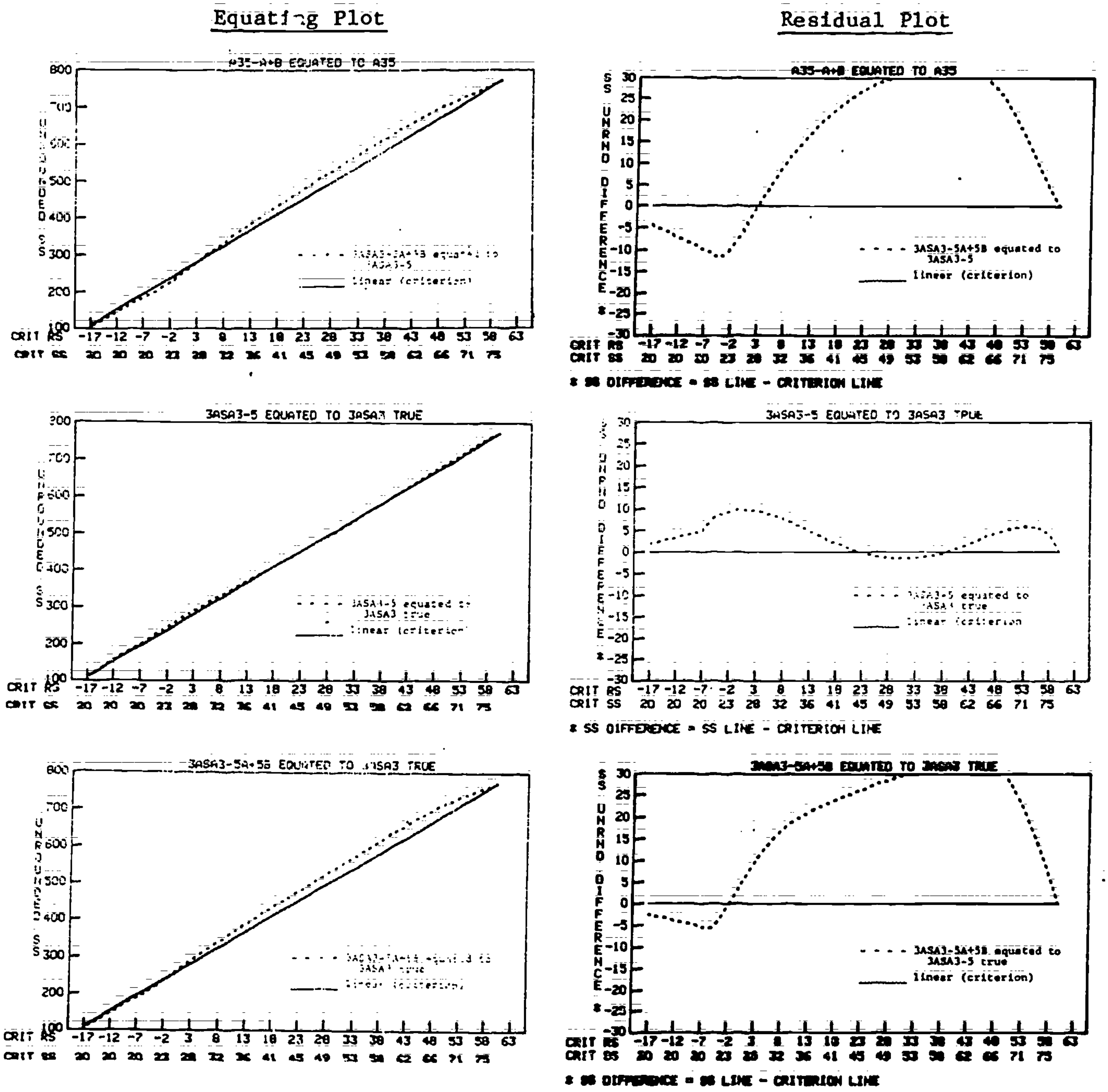

Figure 13. Equating and equating residuals for the second simulation study. Forms 3ASA3-5A+5B equated to 3ASA3-5 (top); Form 3ASA3-5 equated to true form 3ASA3 (middle); form 3ASA3-5A+5B equated to true form 3ASA3 (bottom).

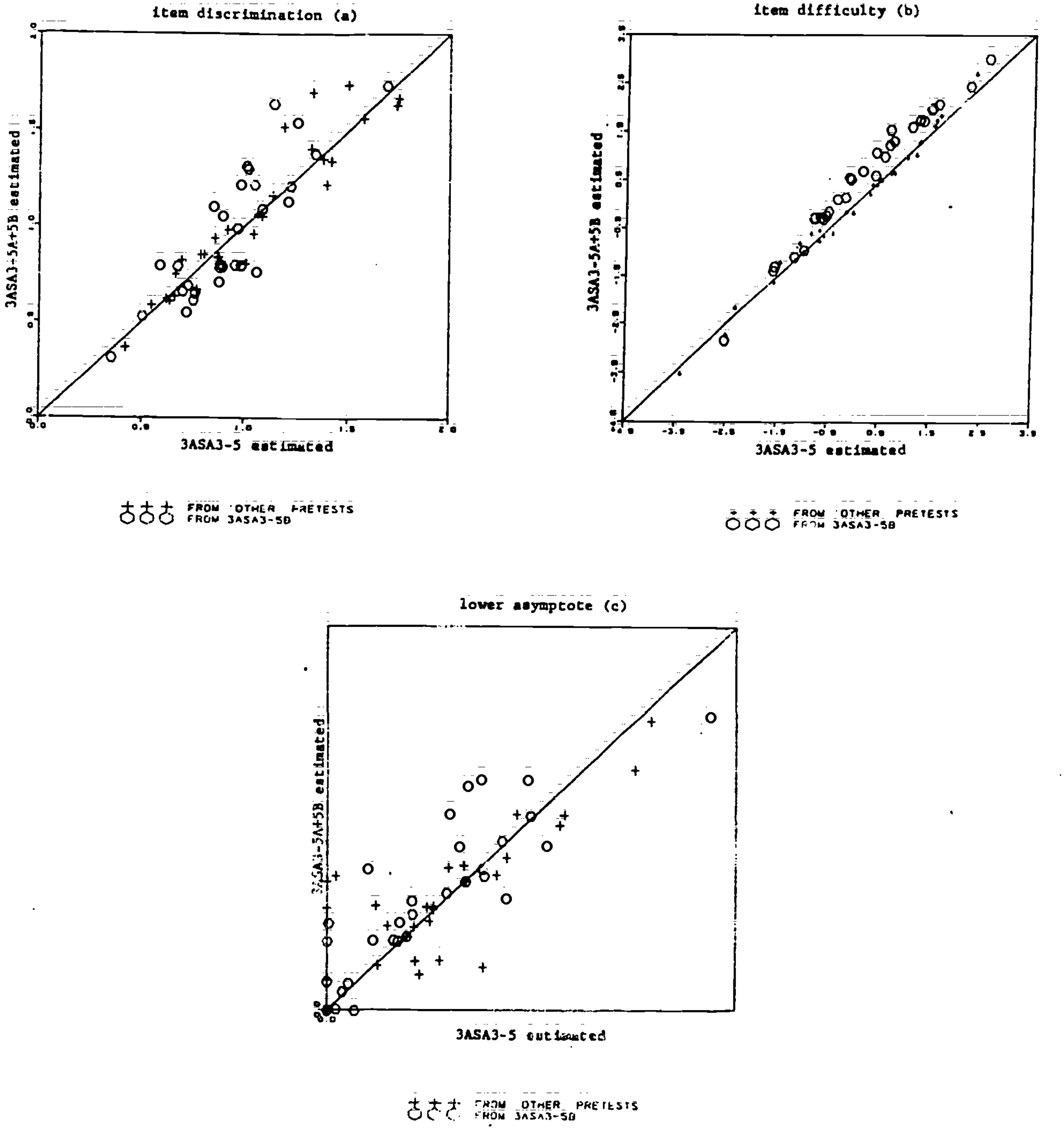

Figure 14. Second simulation study: A comparison of 3ASA3-5A+5B estimated parameters vs. 3ASA3-5 estimated parameters.

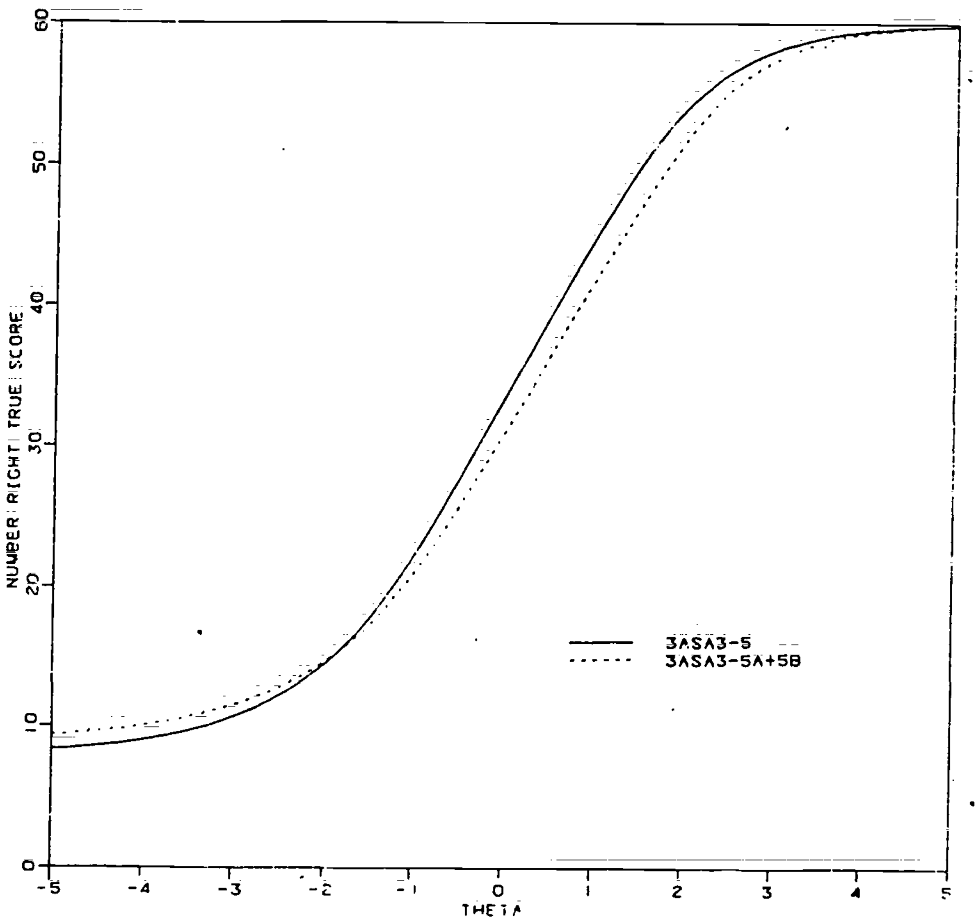

Figure 15. Test characteristic curves for 3ASA3-5 and 3ASA3-5A+5B for the second simulation study.

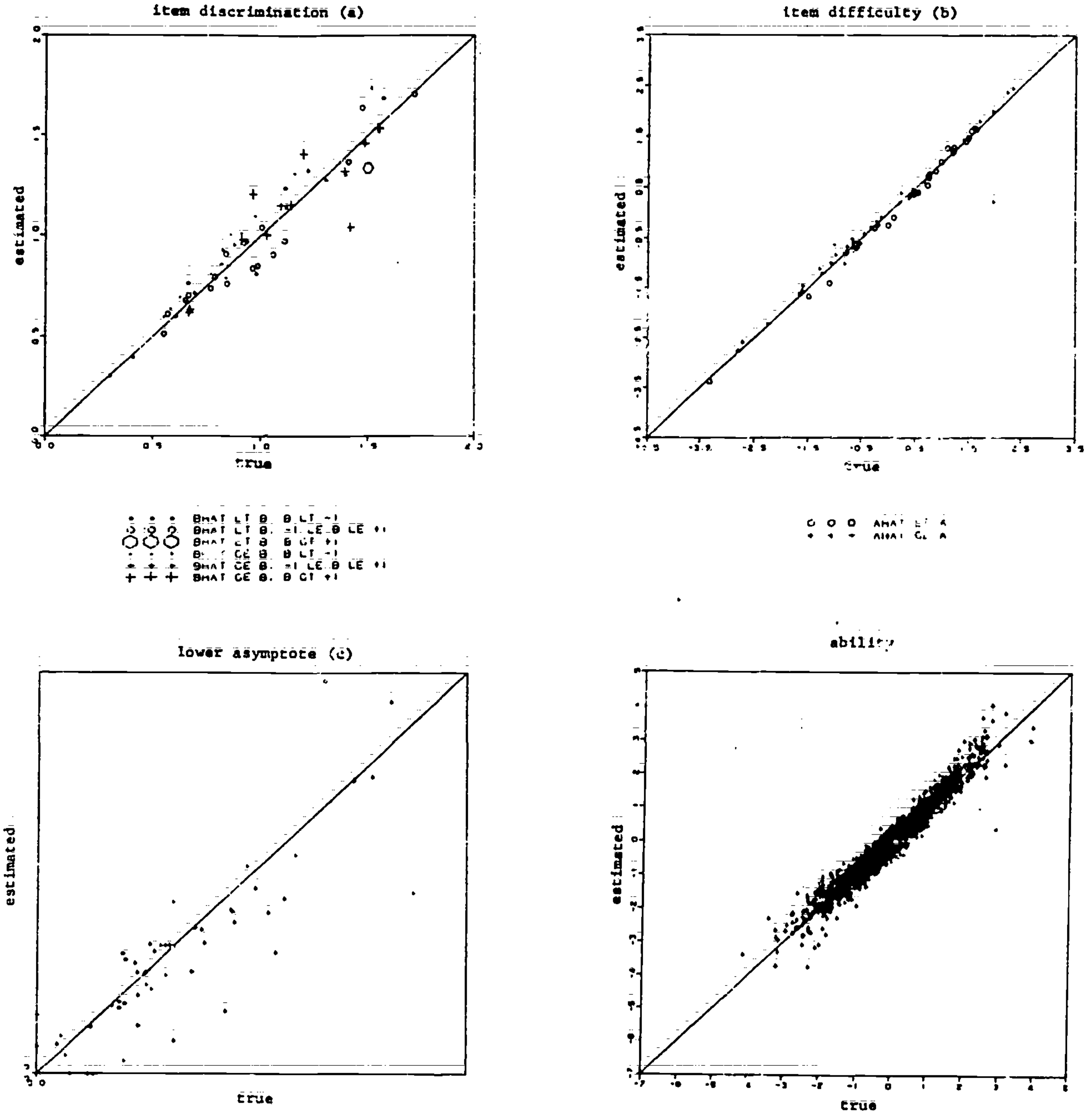

Figure 16: Third simulation study: Parameter estimates for Sample 1 taking form 3ASA3-6 vs. true parameter values.

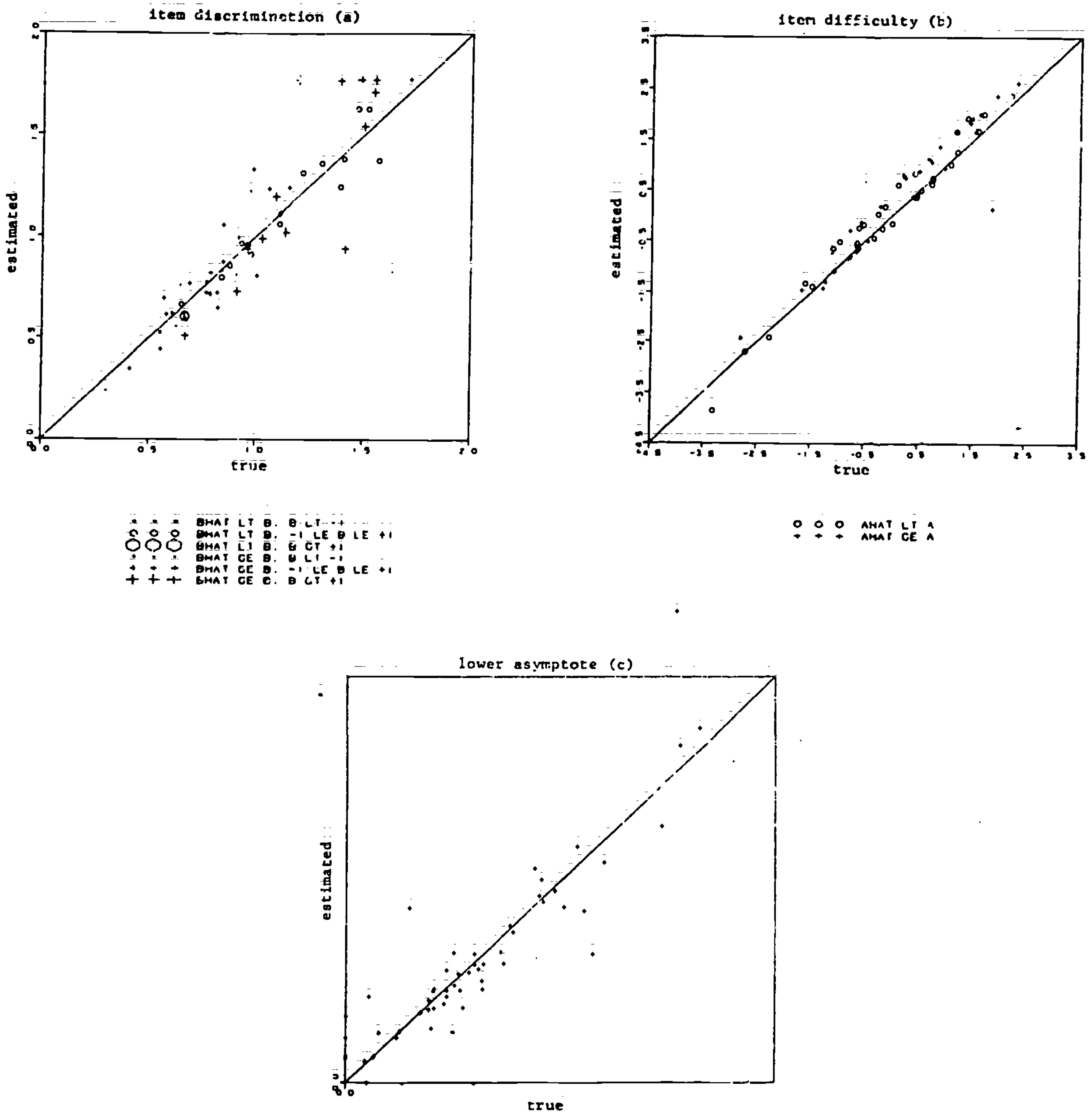

Figure 17. Third simulation study: Parameter estimates for Sample 2 and 3 taking combined form 3ASA3-6A+6B vs. true parameter values.

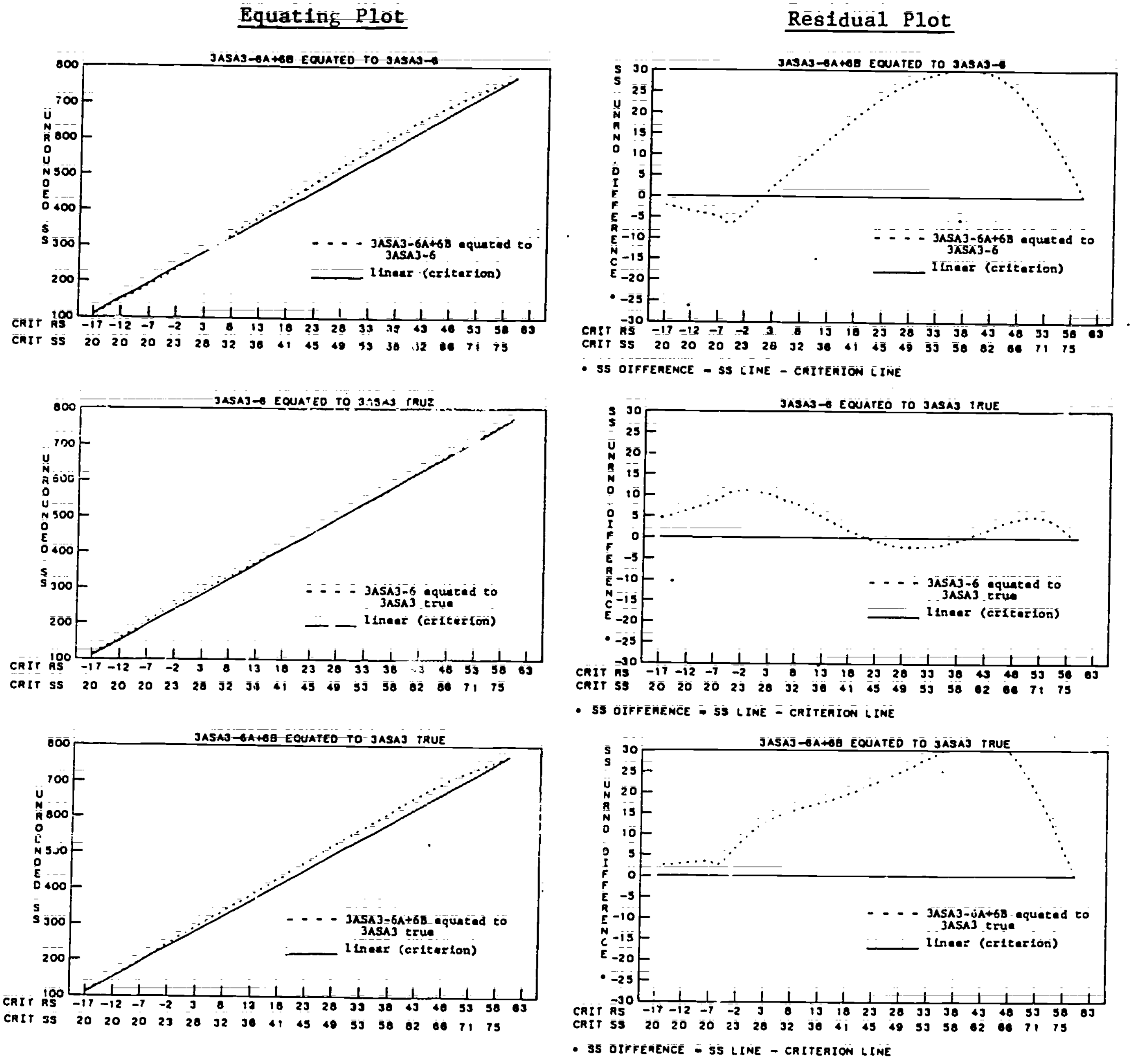

Figure 18. Equating and equating residuals for the third simulation study. Forms 3ASA3-6A+6B equated to 3ASA3-6 (top); Form 3ASA3-6 equated to true form 3ASA3 (middle); form 3ASA3-6A+6B equated to true form 3ASA3 (bottom).

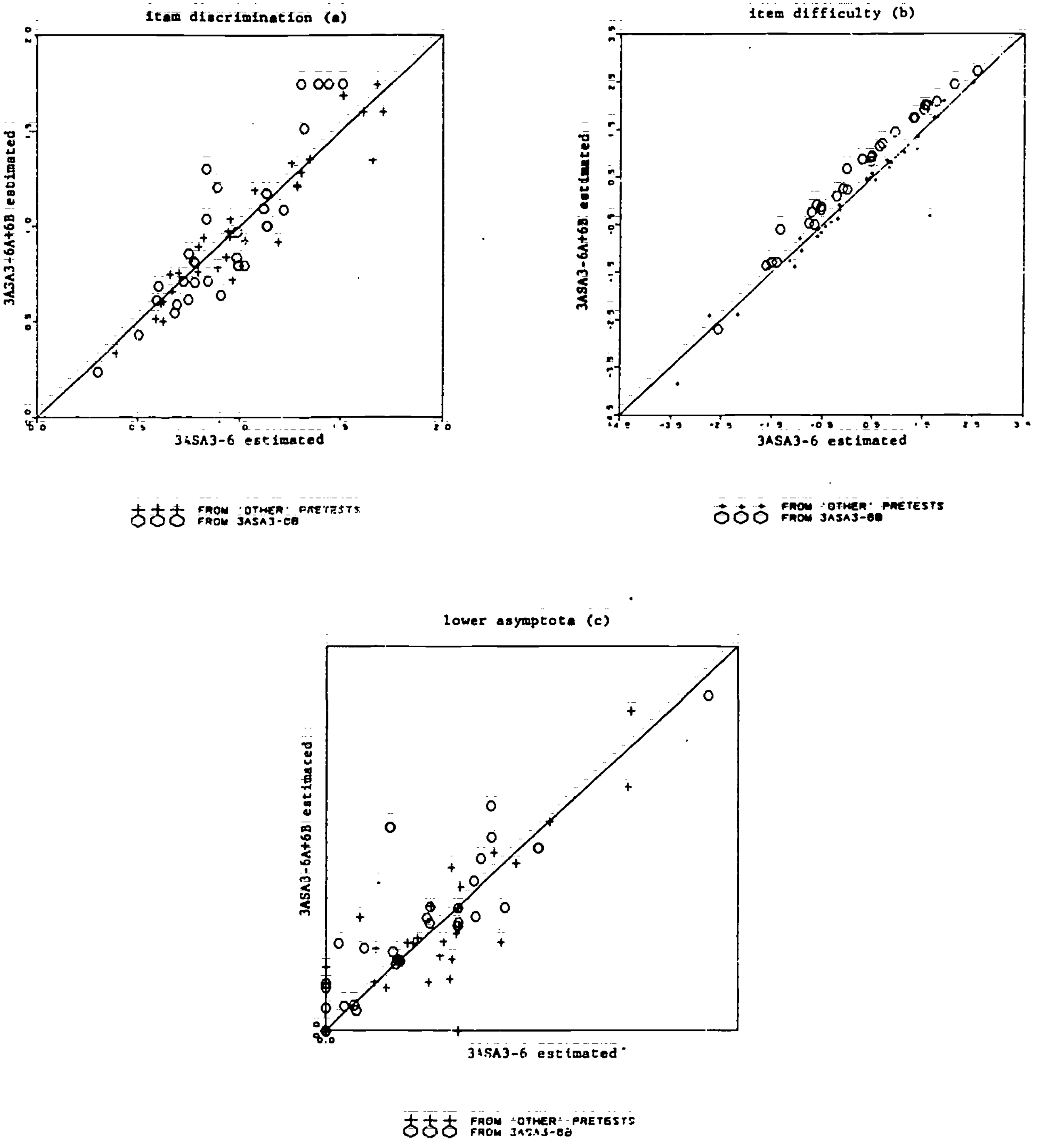

Figure 19. Third simulation study: A comparison of 3ASA3-6A+6B estimated parameters vs. 3ASA3-6 estimated parameters.

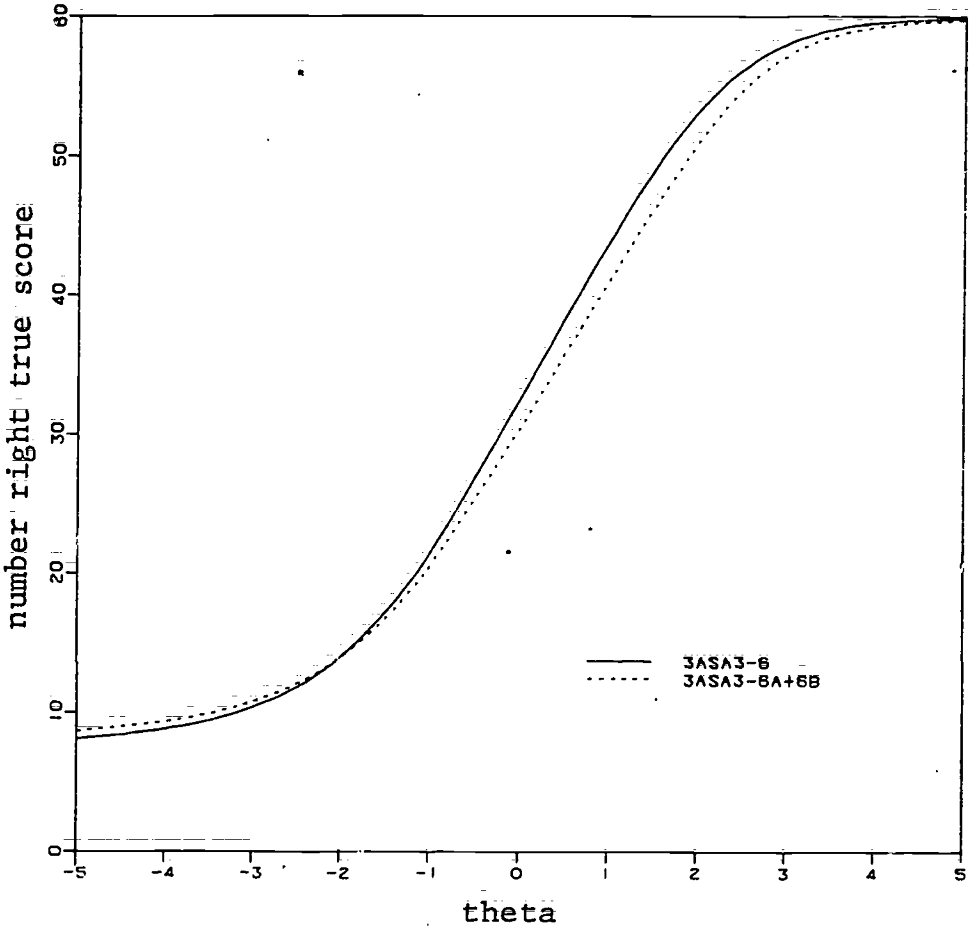

Figure 20. Test characteristic curves for 3ASA3-6 and 3ASA3-6A+6B for the third simulation study.

AN APPENDIX ON SIMILARITIES

The main body of this paper concludes that the third simulation study produces results most consistent with the results for the real data. The third simulation introduced a shift in mean true ability in addition to a particular kind of multidimensionality in the data. This resulted in an overestimation of the item difficulties, and a distorted distribution of estimated abilities. These two results were also seen in the real data. Neither of the other two simulation studies produced these simultaneous results for both item parameter estimates and ability estimates.

It is important to note, however, that just because the third simulation study produced results that resemble real-data results does not imply that the same mechanism, i.e., a decreased mean ability and multidimensionality, necessarily produced the real data. Unfortunately, one can never know what mechanism actually produced the real data. All that can be done is to study as carefully as possible all characteristics of both the real data and the simulated data, looking for further similarities.

In this spirit, the following analysis, suggested by Marilyn Wingersky, was performed. The results provide further evidence for the consistency of simulated and real results.

The Questions to Be Answered

1. Sample 3 from the third simulation study took two blocks of items. On one block, the mean true ability was decreased when compared with other samples. On the other block, mean true ability was decreased and multidimensionality introduced. If abilities were estimated separately from the two sets of items, how would the ability estimates compare with each other?

2. A particular sample of real examinees in the data collection design described in Eignor and Stocking (1986) took two blocks of items. One set was 18 pretest items from pretest form C1613. The other set was a block of 34 items from a section of an operational form, C1. This block of 34 items was combined with additional items not included in the Eignor and Stocking study to produce reported SAT mathematical scores for the sample of examinees. If abilities were estimated separately from the two sets of items, how would the ability estimates compare with each other?

3. Do the two sets of estimated abilities, one for the simulated data and the second for the real data, resemble each other? If so, then the plausibility of the conclusion that both sets of data produced consistent results is strengthened.

The Method, a Caution, and a Standard of Comparison

One mechanism for comparing ability estimates from different sets of items is to examine how well these ability estimates are fit by the estimated response functions for items included in the ability estimate as well as items excluded from the ability estimate. Item-ability regression plots provide a convenient graphical method for making these comparisons. (See Kingston and Dorans (1985), for a detailed explanation of these plots.) The solid curves in the plots used here are the item response functions computed using the estimated item parameters from LOGIST. Each of the different distributions of estimated abilities is grouped identically, and the observed proportions of examinees responding correctly to the item within a particular ability group are plotted with different symbols for each distribution of estimated ability.

For both sets of data, simulated and real, we examined two ability estimates, each based on a sing_ block of items. Necessarily, then, when we examine the item-ability regression for a single item, one ability estimate is based on this and other items in the same block. The other ability estimate is derived from a separate block of items that does not include the item under consideration. Aside from sampling error, we expect on theoretical grounds that the observed proportions for the ability estimates based on separate blocks of items will differ in a systematic way.

In particular, we expect the rough curve formed by the proportions observed for ability estimates that include this item to be steeper than the corresponding rough curve based on ability estimates that exclude this item.

This phenomenon occurs for exactly the same reason that the conventional biserial correlation between item score and total test score is higher when the total test score includes the item under consideration. Lord (1980, p. 33 and p. 40) shows that there is an approximate functional relationship between the biserial correlation and the IRT discrimination parameter under certain restrictive assumptions. If the assumptions are not met, the relationship becomes cruder, but does not disappear.

The rough curves formed by the two sets of observed proportions can be viewed as empirical item response functions. Since the discrimination parameter is a function of the slope of an item response function, we find that the slope is steeper for the empirical curve based on estimated abilities that include the item under consideration.

All of this implies that before we can compare our simulated and real data we need some standard of what is seen when comparing estimated abilities based on different blocks of items under ideal conditions. To produce such a standard, a new set of artificial SAT mathematical data was constructed in which each of 2500 simulees was administered two blocks of items. Each block contained the same 60 items for a total of 120 items per person. The items and simulees were calibrated using LOGIST. The items were then split into the two sets of 60 items and abilities estimated separately using the estimated item parameters for each set of 60 items. Item-ability regressions were then plotted for all items with the two ability estimates plotted with different plotting symbols.

The results are shown for six items in Figures A-1 and A-2. A 'plus' symbol denotes observed proportions from groups of abilities estimated from the first 60 items; a 'hexagon' is used for observed proportions from abilities estimated from the second 60 items. Items 3, 6, and 48 are in the first block of 60; the empirical curve formed by the abilities estimated without these items (hexagons) is less steep than that formed by the abilities estimated from items included in this block (pluses). Items 63, 66, and 92 are in the second block of items. Here the empirical curve formed by the abilities estimated from the first block of items (pluses) is less steep. These six item-ability regressions represent the most noticeable differences between ability estimates based on identical nonoverlapping blocks of items out of the 120 items. They can be used as a standard against which to compare subsequent results.

Insert Figures A-1 and A-2 about here

Results for Simulated Data

Sample 3 in the third simulation was administered 24 items as equating section fn and 30 items as 'pretest' section 3ASA3-6B. Simulees responded to the 24-item block with mean true ability decreased when compared to the other two samples in the third simulation. A particular kind of multidimensionality was introduced into the responses to the 30-item 'pretest' section. Using item parameter estimates from the LOGIST calibration performed in the third simulation, abilities were separately estimated for these two nonoverlapping blocks of items. Item-ability regressions for three items from the 24-item block are shown in Figure A-3 and for three [illegible] from the 'pretest' block in Figure A-4. These particular items were chosen because they show the most discrepancy between the ability estimated.

A 'plus' is used to plot observed proportions from grouped abilities estimated from the 24-item block; a 'hexagon' is used for proportions based on grouped abilities estimated from the 30-item 'pretest' block. In Figure A-3, the observed proportions from grouped abilities estimated from 24-item block (pluses) are reasonably well fit by the estimated item response function. However, the observed proportions from grouped abilities estimated from 'pretest' items (hexagons) are less well fit. A comparison with the standards shown in Figures A-1 and A-2 indicates that this lack of fit is larger than would be expected on the basis of the expected systematic variation alone. The reverse phenomenon is observed for the three 'pretest' items in Figure A-4. Here too the results are larger than the systematic variation shown in the standards of Figures A-1 and A-2. In addition, the observed proportions based on ability estimates from the 24-item block ('pluses') are better fit by the 'pretest' item response functions (Figure A-4) than the observed proportions based on ability estimates from the 'pretest' items (hexagons) are fit by the item response functions in the 24-item block (Figure A-3). That is, the 'pluses' fall closer to the curves in Figure A-4 than the 'hexagons' do in Figure A-3. This is because of the multidimensionality introduced in responses to 'pretest' items.

Insert Figures A-3 and A-4 about here

The conclusions to be drawn from these item-ability regressions are:

1. The abilities estimated from two different sets of items are, in fact, different. This reflects the deliberate modeling of a particular kind of multidimensionality.
2. Given the magnitude of the multidimensionality actually modeled, it is surprising that the abilities estimated from the two different sets of items are as similar as they are.

3. Observed proportions based on abilities estimated from items for which responses were simulated to fit the unidimensional model are fit as well or better by the estimated item response functions than are the observed proportions based on abilities estimated from items for which multidimensionality was introduced.

Results for Real Data

A particular sample of people that were included in a large LOGIST calibration described in Eignor and Stocking (1986) also took two nonoverlapping blocks of items. Eighteen items were pretest items from form C1613; 34 items were items that contributed to the final SAT mathematical score for this sample of examinees and that were included in the large LOGIST calibration. These latter items will be referred to as 'operational.' Not included in the LOGIST calibration were the remaining 25 items that contributed to the final SAT mathematical score for this sample.

Using item parameter estimates from this large LOGIST calibration, abilities were separately estimated the two nonoverlapping blocks of items. Item-ability regressions for three 'operational' items are shown in Figure A-5 and for three pretest items in Figure A-6. These particular items were chosen because they show the most discrepancy between ability estimates.

Insert Figures A-5 and A-6 about here

A 'plus' symbol indicates that ability was estimated from the operational items; a 'hexagon' indicates that ability was estimated from pretest items. In Figure A-5 the observed proportions from grouped abilities estimated from operational items (pluses) are reasonably well fit by the estimated item response function. However, the observed proportions from abilities estimated from the pretest items (hexagons) are less well fit. A comparison with the standards shown ir gures A-1 and A-2 indicates that this lack of fit is larger than would be expected. The reverse phenomenon is observed for the three pretest items in Figure A-6. In addition, the observed proportions based on ability estimates from the operations items (pluses) are fit as well or better by all item response functions shown in Figures A-5 and A-6 than the observed proportions based on abilities estimated from pretest items.

The conclusions that can be drawn from these item-ability regressions are:

1. The abilities estimated from the pretest and operational items are, in fact, different.
2. The nature and magnitude of the differences between the two sets of ability estimates is roughly the same as that seen in the simulated data. This is most easily seen by comparing Figure A-3 with Figure A-5, and Figure A-4 with Figure A-6.

Conclusions from This Investigation

It is clear that the behavior of real examinees studied here was different on pretest items from their behavior on the operational items. The consequences of this different behavior produce results consistent with results for simulated data where a mean shift and a particular type of multidimensionality were introduced. It is also clear that the nature and magnitude of the differences are similar.

The results of this analysis do not prove that the underlying causative mechanisms are the same for both the real and the simulated data. Indeed, there is no analysis that can be performed that will do so. These results do, however, strengthen the assertion that the real and simulated results are consistent with each other.

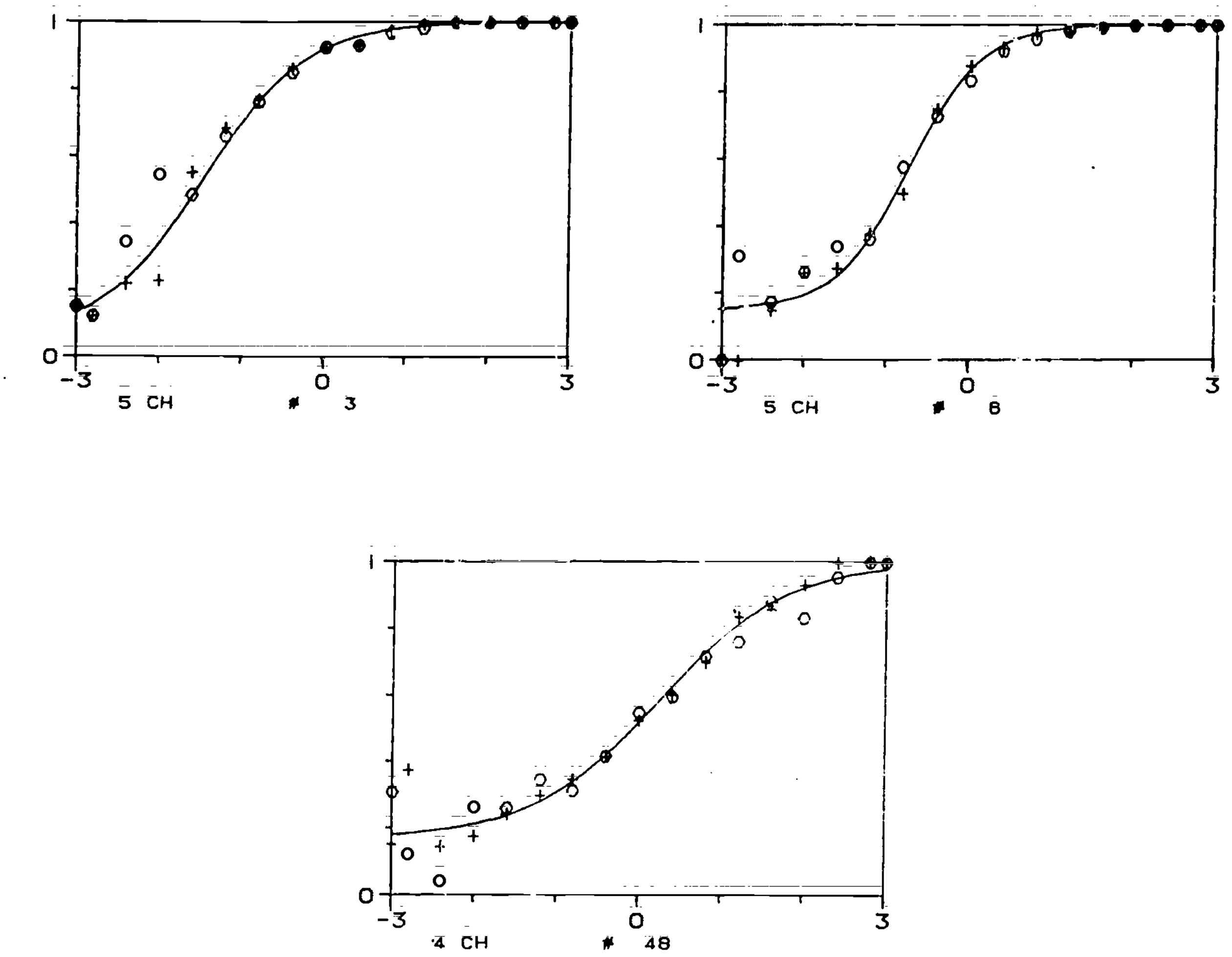

Figure A-1. Item-ability regressions for three items from the first block of 60 items for simulated data where each simulee responded to 120 items. A 'plus' symbolizes observed proportion for abilities estimated on items 1-60; a 'hexagon' symbolizes observed proportions for abilities estimated on items 61-120.

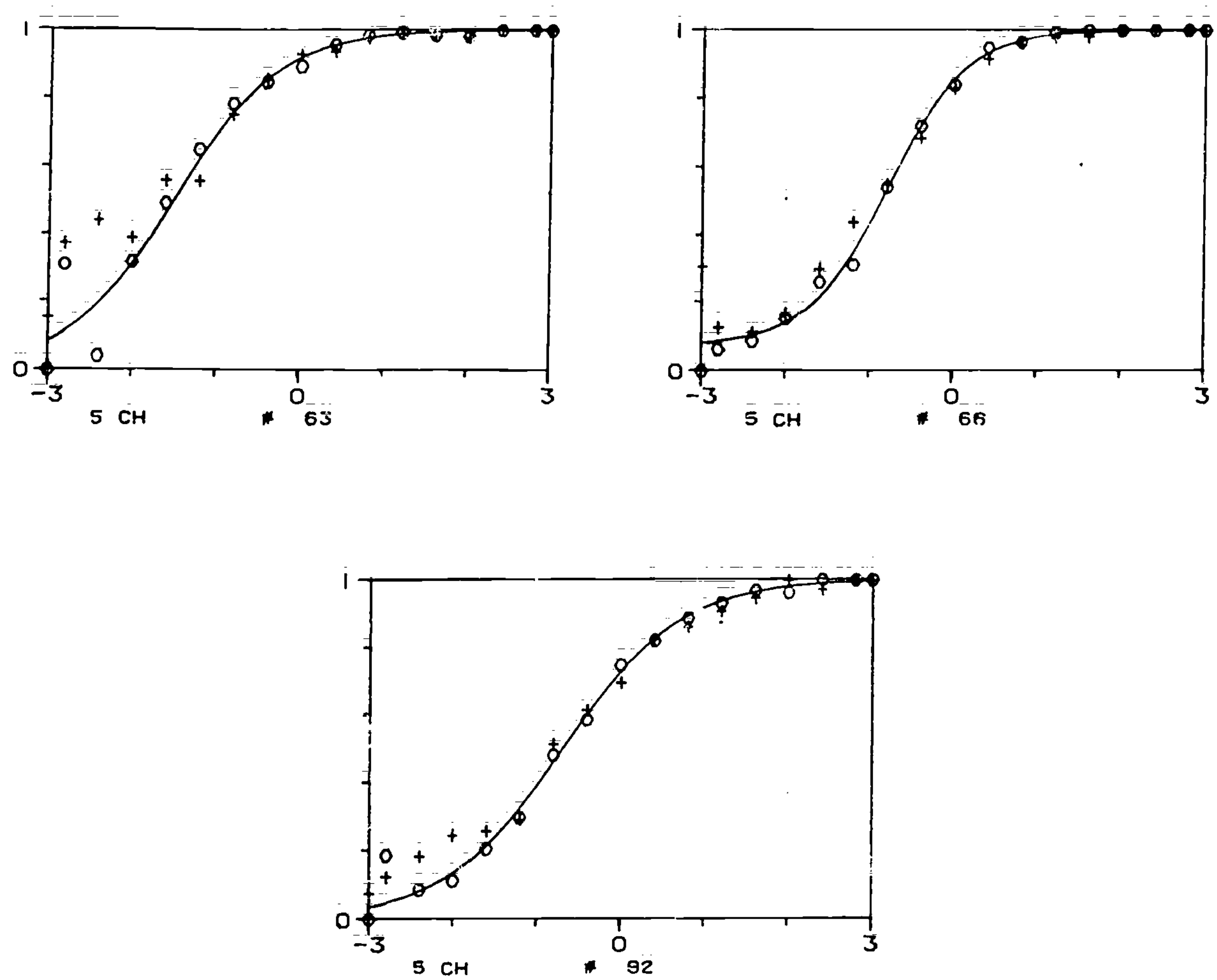

Figure A-2. Item-ability regressions for three items from the second block of 60 items for simulated data where each simulee responded to 120 items. A 'plus' symbolizes observed proportion for abilities estimated on item 1-60; a 'hexagon' symbolizes observed proportions foı abilities estimated on items 61-120.

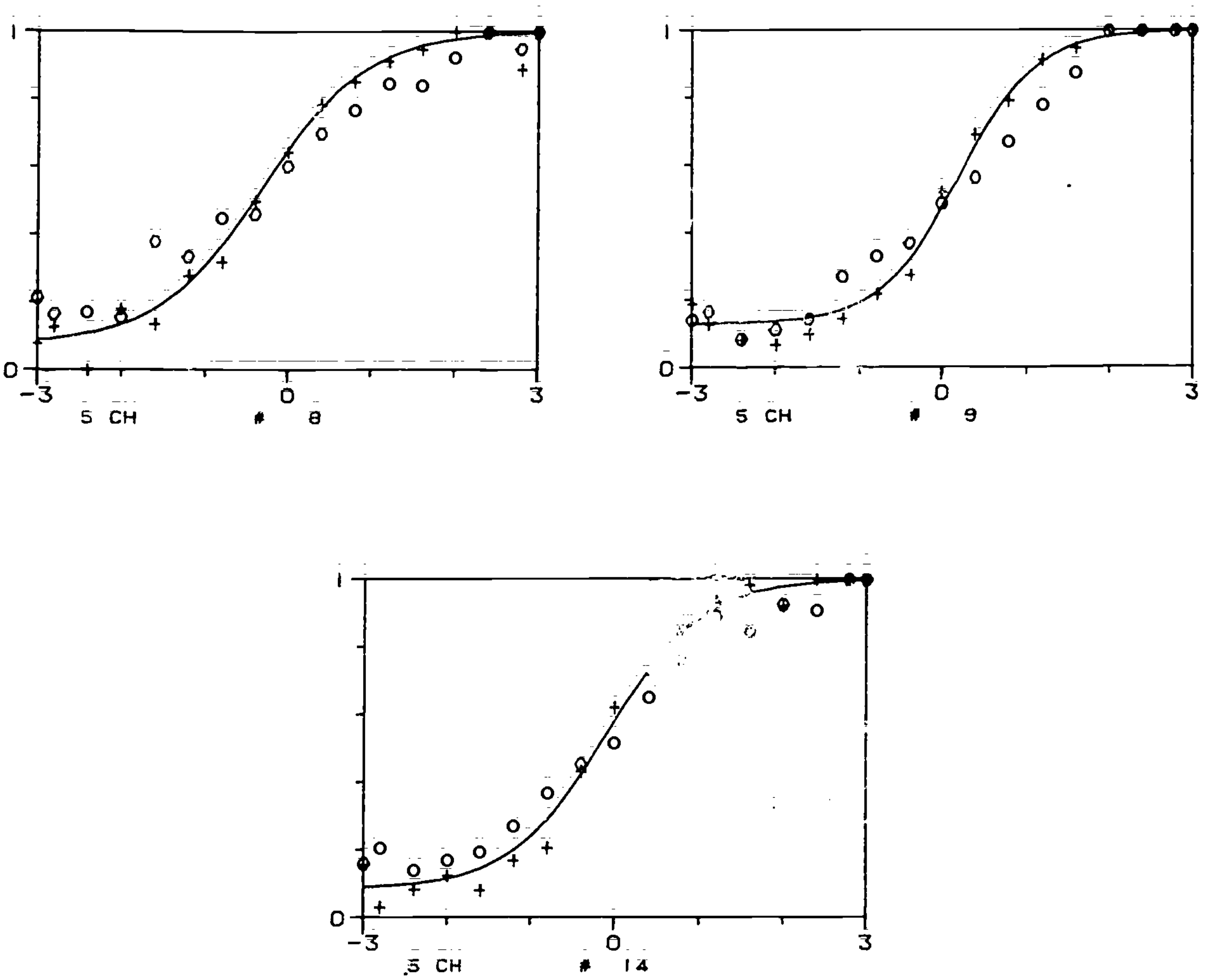

Figure A-3. Item-ability regressions for three items for Sample 3 from third simulation. Responses to these items were simulated with a unidimensional model. A 'plus' symbolizes observed proportions for abilities estimated on the block of items for which responses were unidimensional; a 'hexagon' symbolizes observed proportions for abilities estimated on the block of items for which mutidimensionality was introduced.

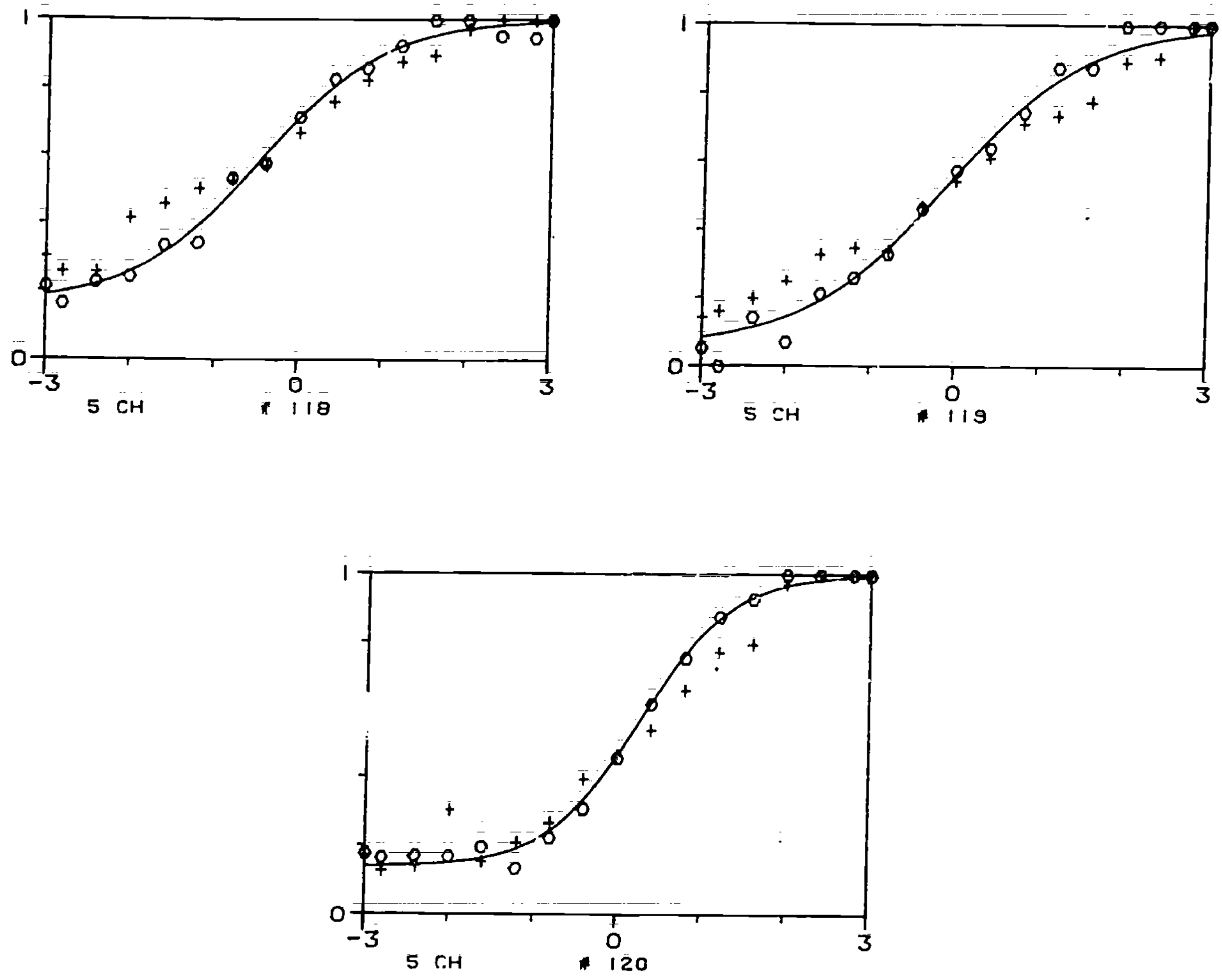

Figure A-4. Item-abil'ty regressions for three items for Sample 3 from third simulation. Responses to these items were simulated with a particular unidimensional model. A 'plus' symbolizes observed proportions for abilities estimated on the block of items for which responses were unidimensional; a 'hex[illegible]n' symbolizes observed proportions for abilities estimated [illegible]ne block of items for which mutidimensionality was introduced.

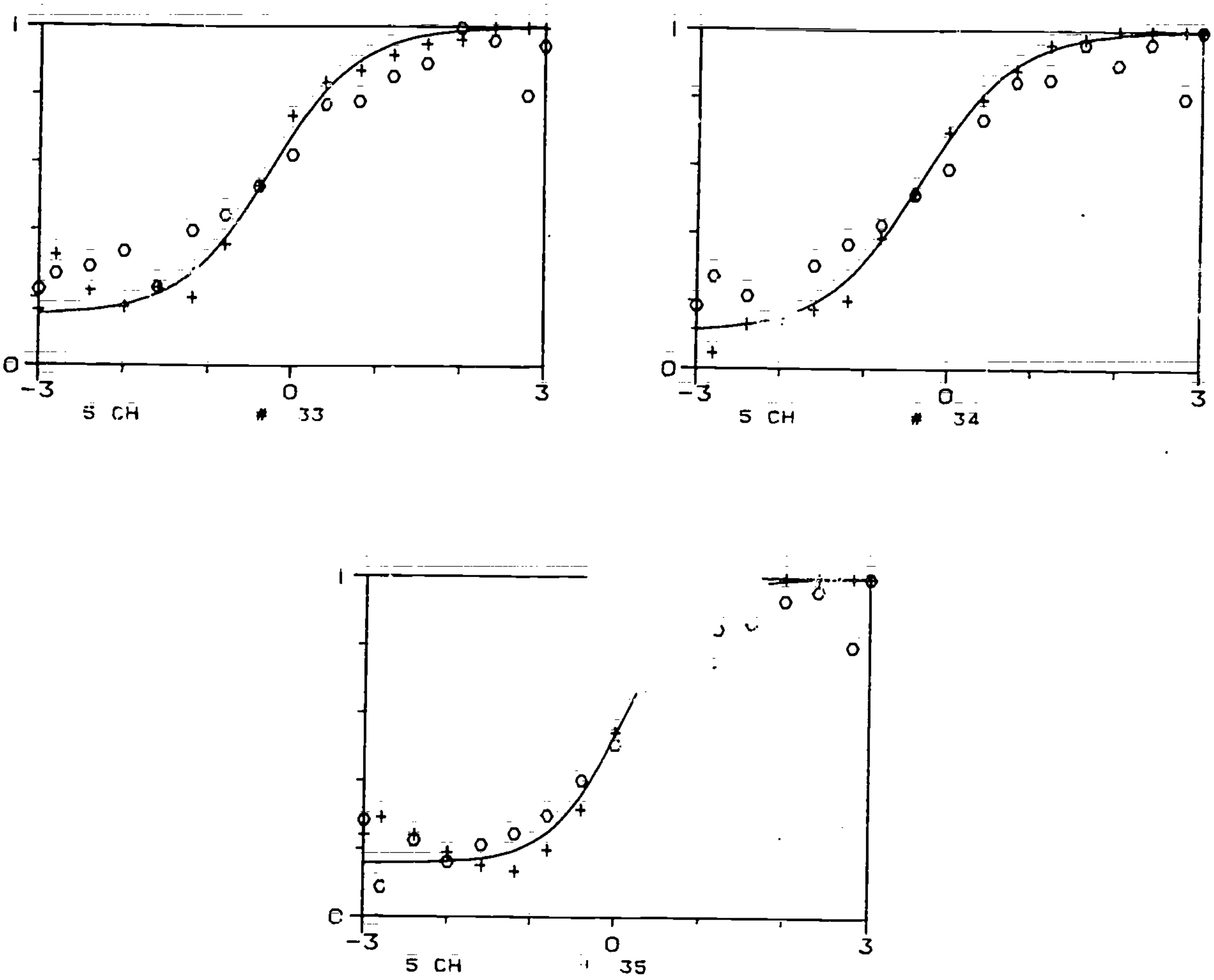

Figure A-5. Item-ability regressions for real data, sample 10 from Eignor and Stocking (1986). These three items are from the block of operational items. A 'plus' symbolizes observed proportions for abilities estimated on operational items; a 'hexagon' symbolizes observed proportions for abilities estimated on pretest items.

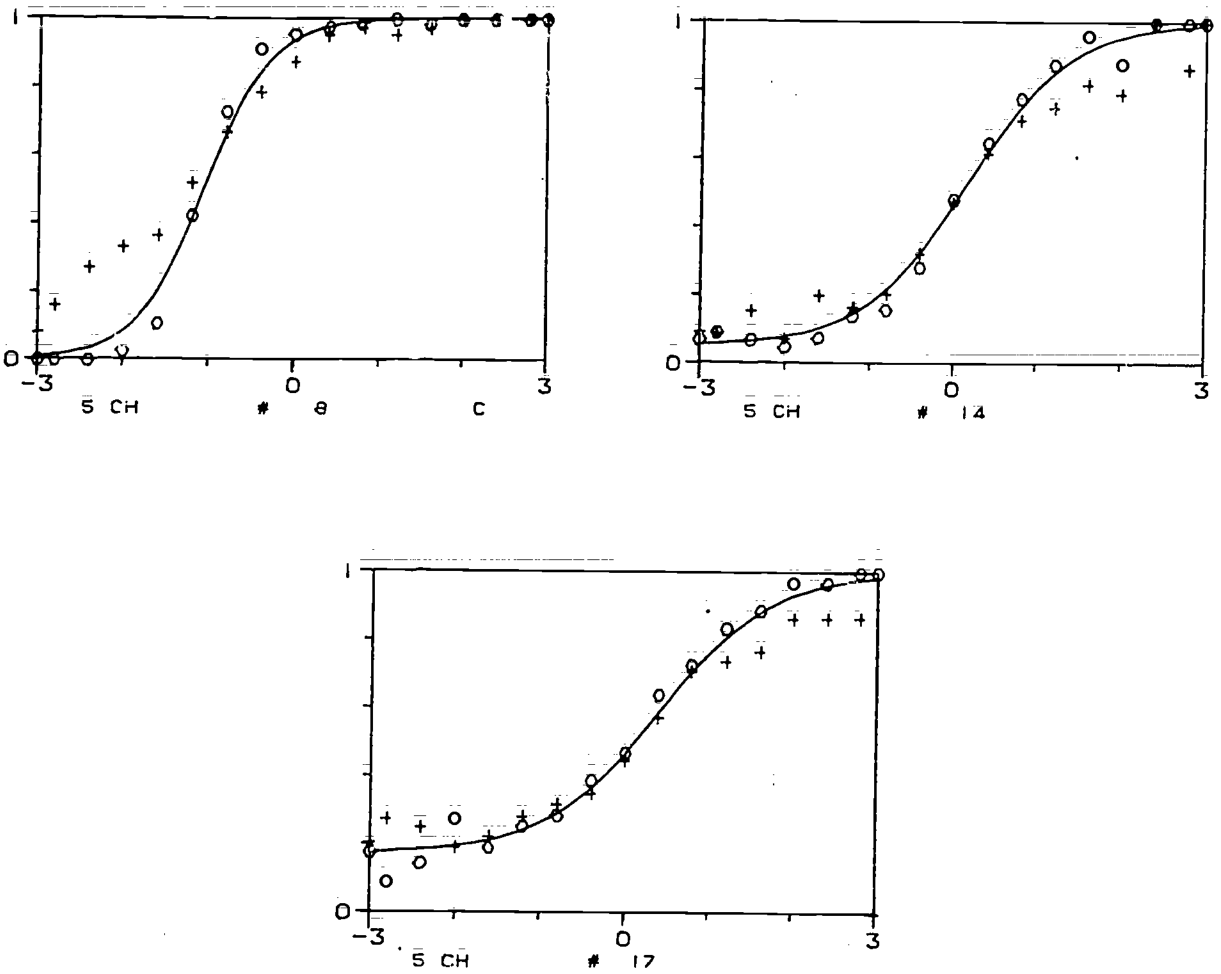

Figure A-6. Item-ability regressions for real data, sample 10 from Eignor and Stocking (1986). These three items are from the block of pretest items. A 'plus' symbolizes observed proportions for abilities estimated on operational items; a 'hexagon' symbolizes observed proportions for abilities estimated on pretest items.