GEORGIA INSTITUTE OF TECHNOLOGY OFFICE OF CONTRACT ADMINISTRATION

PROJECT ADMINISTRATION DATA SHEET

☒ ORIGINAL □ REVISION NO. ______

Project No. M-50-601 GTRI/XXX DATE 4/20/83

Project Director: Dr. Charles K. Parsons School/XXX Management

Sponsor: AFOSR; Bolling AFB, D.C. 20332

Type Agreement: Grant No. AFOSR-83-0172

Award Period: From 7/1/83 To 6/30/84 (Performance) ______ (Reports)

Sponsor Amount: Total Estimated: $ 11,991 Funded: $ 11,991

Cost Sharing Amount: $ ______ Cost Sharing No: ______

Title: Multidimensional Tests and Item Bias; A Proposal for a Monte Carlo Study

ADMINISTRATIVE DATA

OCA Contact John W. Burdette x-4820

1) Sponsor Technical Contact:

Major Carl Edward Oliver
AFOSR/NM
Building 410
Bolling AFB, DC 20332

Phone: (202) 767-5025

2) Sponsor Admin/Contractual Matters:

Hugh M. McElroy
AFOSR/PKZ
Building 410
Bolling AFB, DC 20332

Phone: (202) 767-4952

Defense Priority Rating: N/A

Military Security Classification: N/A
(or) Company/Industrial Proprietary: N/A

RESTRICTIONS

See Attached AFOSR Supplemental Information Sheet for Additional Requirements.

Travel: Foreign travel must have prior approval – Contact OCA in each case. Domestic travel requires sponsor approval where total will exceed greater of $500 or 125% of approved proposal budget category.

Equipment: Title vests with N/A; None proposed.

COMMENTS:

COPIES TO:

Research Administrative Network
Research Property Management
Accounting
Procurement/EES Supply Services

Research Security Services
~~Reports Coordinator (OCA)~~
GTRI
Library

Research Communications (2)
Project File
Other Parsons
Other ______

GEORGIA INSTITUTE OF TECHNOLOGY　　　　OFFICE OF CONTRACT ADMINISTRATION

SPONSORED PROJECT TERMINATION/CLOSEOUT SHEET

Date 3/25/86

Project No. M-50-601　　　　School/XXX Management

Includes Subproject No.(s) None Indicated

Project Director(s) C. Parsons　　　　GTRC / ~~GIT~~

Sponsor AFOSR - Bolling, AFB, D.C.

Title Multidimensional Tests and Items Bias: A Proposal for a Monte Carlo Study

Effective Completion Date: 8/30/84 (Performance) 8/22/84 (Reports)

Grant/Contract Closeout Actions Remaining:

☐ None

☒ Final Invoice or Final Fiscal Report

☐ Closing Documents

☐ Final Report of Inventions

☐ Govt. Property Inventory & Related Certificate

☐ Classified Material Certificate

☐ Other ____________________

:ontinues Project No. None Indicated　　　　Continued by Project No. None Indicated

:OPIES TO:

'roject Director	Library
:esearch Administrative Network	GTRC
esearch Property Management	Research Communications (2)
ccounting	Project File
·ocurement/GTRI Supply Services	Other A. Jones
esearch Security Services	M. Heyser
:ports Coordinator (OCA)	R. Embry
:gal Services	

1983-1984 Research Initiation in Science and Engineering Program

Sponsored by

Air Force Office of Scientific Research

Conducted by

Southeastern Center for Electrical Engineering Education

Final Report Draft

A Monte Carlo Study of Item Bias Detection in Multidimensional Tests

Prepared by: Charles K. Parsons, Ph.D.

Academic Rank: Assistant Professor

Department and University: College of Management
Georgia Institute of Technology

Research Location: Georgia Institute of Technology

Date: August 19, 1984

Contract Number: AFOSR-83-0172

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER M-50-601	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Monte Carlo Study of Item Bias Detection in Multidimensional Tests		5. TYPE OF REPORT & PERIOD COVERED 7-1-83 to 6-30-84 Final Report (Draft)
		6. PERFORMING ORG. REPORT NUMBER M-50-601
7. AUTHOR(s) Charles K. Parsons		8. CONTRACT OR GRANT NUMBER(s) AFOSR-83-0172
9. PERFORMING ORGANIZATION NAME AND ADDRESS College of Management Georgia Institute of Technology Atlanta, Georgia 30332		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS AFOSR/NM Building 410 Bolling AFB, DC 20332		12. REPORT DATE August 20, 1984
		13. NUMBER OF PAGES 48
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Item Bias, Test Bias, Multidimensional Test, Item Response Theory, Item Characteristic Curve, Item Parameters

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study concerns the effects of test multidimensionality on recommended item bias statistics. Simulation data samples (N=1000 each) on a 50 item test were generated using a factor model described and used by Drasgow and Parsons (1983) and Parsons (1982). Subpopulation differences on common factors led to item bias that was identified to some extent by both chi-square and item response theory (2 parameter logistic) bias indices. The signed indices were especially effective in distinguishing biased items from

Abstract

This study concerns the effects of test multidimensionality on recommended item bias statistics. Simulation data samples (N=1,000 each) on a 50 item test were generated using a factor model described and used by Drasgow and Parsons (1983) and Parsons (1982). Subpopulation differences on common factors led to item bias that was identified to some extent by both chi-square and item response theory (2 parameter logistic curve) bias indices. The signed indices were especially effective in distinguishing biased items from unbiased items. However, the use of either the signed chi-square or signed IRT index in multidimensional data clearly requires an _a priori_ knowledge of which subpopulation is at a disadvantage. This rather unexpected finding suggests further study of the properties of signed indices as well as a reevaluation of previous simulation research that has appeared to support their validity.

A MONTE CARLO STUDY OF ITEM BIAS DETECTION IN MULTIDIMENSIONAL TESTS

May, 1984 Draft

Mental tests have had a controversial history. Persistent differences between racial groups on standardized aptitude test scores have suggested the potential for unfair discrimination against members of different racial and ethnic subpopulations. Because many occupational and educational opportunities are affected by test scores, the issue of test bias has consequences for many people in our society.

The study of test bias has intensified since the passage of the Civil Rights Act of 1964. Arvey (1979) provides a review of much of the research in this area. Some researchers have preferred to concentrate on item rather than test bias; logically inferring that a biased test must contain biased items. This approach allows the possibility that some items on a test are biased while others are not.

More than a dozen statistical techniques have been proposed for detecting biased items. Various techniques have been studied theoretically and empirically with real and simulated data (see Hulin, Drasgow & Parsons, 1983; Berk, 1982).

There appears to be a preference for techniques based on a latent trait or item response theory (IRT) because sample estimates of population item parameters are invariant. This advantage occurs because, when the IRT model is valid, item parameters are invariant with respect to subpopulation ability distributions. Except for sampling error, any item should have identical IRT item parameter estimates (within a linear transformation) in two or more subpopulations, regardless of ability distribution. Usually, the specific IRT model, has been the three parameter logistic model (Birnbaum, 1968) and has been referred to as Item Characteristic Curve-3 or ICC-3 though all IRT models have the parameter invariance property.

The main drawback to ICC-3 is the costly estimation of model parameters and hence bias statistics. Therefore, some previous empirical research has focused on identifying less costly techniques that converge with IRT methods.

Another aspect of item bias that deserves attention is the robustness of item bias statistics when assumptions of the IRT model are violated. For instance, the assumption of unidimensionality is likely to be violated on tests that are developed to predict external job or educational performance criteria. Much of the concern about unidimensionality is how to define it operationally (McDonald, 1981). Recently, Drasgow and Parsons (1983) and Parsons (1982) have shown that unidimensional IRT models can be applied and interpreted in multidimensional tests. These latter studies will be described later. However, the consequences of multidimensionality for item bias statistics are unknown. This paper will proceed with a definition of item bias as well as a review of some studies that have compared the relative effectiveness of various statistical indices of bias. Then, a brief description of the Drasgow and Parsons (1983) methods for generating multidimensional data for item response simulations will be given. Finally, four cases of test multidimensionality and item bias will be described from which sample data will subsequently be generated and analyzed for the effectiveness of item bias statistics.

Defining Item Bias

Item bias is present when individuals from different subpopulations with the same amount of a latent ability tend to have different probabilities of responding correctly to an item. One cause of item bias would be the use of subpopulation specific information in an item (e.g., information not relevant to the construct or criterion of interest). In this case, members of one subpopulation have an unfair advantage and would have a higher probability of a correct answer. Note that biasing factors should be considered as continuous variables and not as distinguishing characteristics between subpopulations. Clearly, not all members of one group will have exposure to some culturally specific information, while all members of another group will have been denied exposure to the information. Just as mean differences on test score distributions are associated with mean differences but overlapping distributions on primary abilities, we will associate possible mean differences on secondary attributes with overlapping distributions.

Item Bias Statistics

There are a variety of statistics that have been suggested for identifying item bias. Rather than review them here, the reader is referred to an article by Ironson and Subkoviak (1979) or books by Hulin, Drasgow and Parsons (1983) and Beck (1982) for reviews.

Most previous studies have used an IRT index based on the logistic model and a Chi-Square index based on total score intervals. The general IRT approach will be described first. In item response theory, the probability of a correct response to an item is a function of one or more item parameters and the latent ability of the examinee. In the case of multiple choice tests, the logistic model could be the three parameter logistic model or:

$$P_i(\Theta) = C_i + (1-C_i)/\left[1+\exp(-a_iD(\Theta-b_i))\right] \qquad (1)$$

where $P_i(\Theta)$ is the probability of a correct response on item i for a given level of Θ (latent ability), a_i is item discrimination, b_i is item difficulty, D is a scaling factor usually set to 1.702, and C_i is the pseudo-guessing parameter that reflects the probability of a correct response at very low levels of Θ. Other possible logistic models are the one parameter model (Rasch, 1960; Wright, 1977) or the two parameter logistic model (Lord and Novick, 1968). Choosing among the models depends on the user's goals. Actual responding to multiple choice questions would seem best described by the three parameter model above, but others have argued that the mathematical properties of the one parameter model make it better in many cases. In the current study, the author chose to use the two parameter model because it best met the demands of the current study. This will be elaborated upon later in this paper.

Regardless of the model chosen, item bias can be defined in a number of ways within the IRT framework. First, one can test for the difference between parameters estimated in two or more subpopulations (Lord, 1980). More common recently has been the computation of differences between two item characteristic curves estimated in two subpopulations. Suppose that $P_1(\Theta)$ represents the function relating Θ to the probability of a correct response for an item in Subpopulation 1 and $P_2(\Theta)$ represents the corresponding function in Subpopulation 2. Then, the area between these two item characteristic curves could be approximated by:

$$\text{ALGICC} = \sum_{j=1}^{301} \left(P_1(\Theta_j) - P_2(\Theta_j)\right) \times .02 \qquad (2)$$

$P_1(\Theta_j)$ is the probability of a correct response at Θ level j in subpopulation 1 and $P_2(\Theta_j)$ is the probability of a correct response at Θ level j in group 2. The value .02 is the width of the theta interval. Note that this statistic would tend to reflect an average difference between curves with differences in different directions cancelling each other out. This approach will be called the Algebraic Sum approach and will be referred to as ALGICC in this paper.

A statistic that does not average the differences between curves is the Absolute Sum approach and is computed as:

$$\text{ABSICC} = \sum_{j=1}^{301} \text{ABS}\left(P_1(\Theta_j) - P_2(\Theta_j)\right) \times .02 \qquad (3)$$

where ABS is the absolute value function. In this case, regardless of the direction of the difference, it is still added to the bias statistic.

The chi-square approaches are conceptually similar to the IRT approaches but without the theoretical elegance. Scheunemann (1979) suggested that an investigator could sort examinees from each subpopulation into five intervals based on total test score. For each subpopulation in each interval, the proportion of correct responses is computed and compared to the expected proportion. These proportions are then combined over total score intervals to compute a chi-square or:

$$\text{ABSCHI} = \sum_{j=1}^{5} N_{1j} \frac{(P_{1j} - P_{Tj})^2}{P_{Tj}(1 - P_{Tj})} + N_{2j} \frac{(P_{2j} - P_{Tj})^2}{P_{Tj}(1 - P_{Tj})} \qquad (4)$$

where P_{1j} represents the total correct on an item and N_{1j} is the number of examinees for subpopulation 1 in total score interval j, while P_{2j} is the proportion and sample size for subpopulation 2. P_{Tj} and N_{Tj}represent the proportion and sample size for both samples combined. In this case, as in ABSICC, the statistic does not reflect direction of difference. In order to include direction, an algebraic sum is computed as

$$\text{ALGCHI} = \sum_{j=1}^{5} S_jN_{1j} \frac{(P_{1j} - P_{Tj})^2}{P_{Tj}(1 - P_{Tj})} + S_jN_{2j} \frac{(P_{2j} - P_{Tj})^2}{P_{Tj}(1 - P_{Tj})} \qquad (5)$$

where $S_j = 1$ if $P_{1j} < P_{2j}$, $S_j = -1$ if $P_{1j} > P_{2j}$, and $S_j = 0$ if $P_{1j} = P_{2j}$ and will be referred to as ALGCHI.

Several recent studies have compared the effectiveness of these statistics. Rudner, Getson, and Knight (1980) compared two transformed item difficulty approaches, three IRT methods, and two chi-square approaches in simulated item response data sets generated from 112 different combinations of test conditions. Item responses were generated on the basis of the three parameter logistic model with varying degrees of bias built into the responses from different groups. Of particular interest in this study was the finding that the five interval signed chi-square index (ALGCHI) was found to perform quite well. The authors concluded that "with five total score intervals, the chi-square technique was found to be as effective as the three parameter item characteristic curve theory technique, under most of the investigated conditions" (Rudner et al., 1980, p. 8). It should be noted that across all 112 test conditions, the three parameter logistic index correlated .80 with generated bias. While the 5 score interval Chi-Square index correlated .73 with generated bias. Other results for their study showed that correlations with generated bias ranged from a low of .55 for the one parameter IRT model to .68 for a transformed item difficulty index.

Shepard, Camilli and Averill (1981) examined the performance of various indices in real data. They chose a total of 16 item bias indices which were variations of the transformed item difficulty method, the item discrimination method, the three parameter IRT method, the one parameter IRT method, and the chi-square method. They studied convergence of these methods in samples of 490 black, 551 Chicano, and 552 white pupils in the fourth, fifth, and sixth grades. The dimensionality of the utilized Lorge-Thorndike verbal and nonverbal tests was described on the basis of the size of the first principal factor relative to the other factors. Variance accounted for by the first factor ranged from 17.5% for the black sample to 19.7% for the white sample. The size of subsequent factors dropped to about 5%. The authors interpreted this as supporting the contention of one general factor in the test.

Though this study could not determine correlations with true bias because the data were real and not based on a simulation model, a factor analysis of correlations between the 16 indices indicated that the full unsigned chi-square (ABSCHI) method loads on the same factor as the unsigned ICC-3 index (ABSICC). Also, the signed ICC-3 (ALGICC) and signed chi-square

(ALGICC) methods both loaded on another factor. Based on this and other results, the authors conclude that "it may be safe to recommend that the signed full chi-square technique could be used as the best substitute for the ICC-3 _b_ differences and signed area" (Shepard et al., 1981).

Multidimensionality and Item Bias

In studies of real data (e.g., Shepard et al., 1981) observed differences in various item statistics between groups could also be due to a lack of fit of the model to the data. One major concern has been that the data will most likely be multidimensional. The frequently used unidimensional IRT models (due to availability) can be fit to multidimensional data (Drasgow and Parsons, 1983; Reckase, 1979) but the conceptual meaning of item bias as well as the effect on computed bias statistics has not been given much attention. Researchers may regard multidimensionality as "spurious" bias rather than true bias. Spurious bias is that which tends to invalidate statistical indices of bias by including lack-of-fit and sampling error effects. For example, Shepard et al. (1981) state "that when parameter estimates are not linearly related across groups this may be due to a lack of model fit rather than bias. For example, a violation of unidimensionality assumption could be detected as bias when group differences are not the same across traits -- an effect that is still consistent with the interpretation of bias." Shepard et al. go on to say "If a test were multidimensional, differential differences in group abilities across factors would appear as bias." These authors seem to be implying that item bias requires that the test data be very well fit by the unidimensional IRT model. On the other hand, Linn, Levine, Hastings, and Wardrop (1982) consider multidimensionality a form of bias because "it can lead to apparent differences in the primary ability when, in fact, there are no such differences." Before forming research questions concerning multidimensionality and item bias, it is constructive to review the recent research on multidimensionality and IRT.

Multidimensionality and Unidimensional IRT

Of all possible IRT models, unidimensional IRT models have had, by far, the greatest application among IRT testing practitioners and theorists. In constructing tests, practitioners are advised to retain items that conform to the unidimensionality assumption. Only recently have researchers begun to investigate the implications of violating the assumption.

Reckase (1979) provided an interesting demonstration of what happens when unidimensional IRT models are fit to tests with one or several major factors. It appeared that a test with one large first factor could be fit quite well in spite of the presence of other factors. On the other hand, when two major factors were present in the test, the unidimensional model appeared to fit only items that loaded highly on one of the two factors.

In a Monte Carlo study, Drasgow and Parsons (1983) expanded the scope of Reckase's (1979) study by including finer gradations of the size of the major factor. They found that for both the two and three parameter logistic models, the unidimensional IRT model fit test data that had a surprisingly high degree of heterogeneity among items. Drasgow and Parsons suggested that strict tests of unidimensionality might reject items and tests that they had found to be estimated satisfactorily.

Parsons (1982), using the same Monte Carlo methodology, found that under the condition of oblique factors, additional factors in the test actually improved the fit of the IRT model. He pointed out that as the number of correlated factors approaches the number of items, the independent variance of each factor becomes small relative to the variance shared with the other oblique factors. This caused the test to conform better to an IRT model with only one dimension. Note that the above results were obtained with oblique, not orthogonal, factors.

The Simulation Model

The basic model of multidimensional tests used in this and previous studies (Drasgow and Parsons, 1983; Parsons, 1982) had a general latent ability that affected all items on the test and more specific abilities that were each limited to a subset of items on the test.

Multidimensional data structures with potentially biasing specific ability factors can be represented by the common factor model (Thurstone, 1947) but further transformation is necessary to adequately model dichotomously scored item responses. The common factor model represents continuous observed variables, x, as weighted linear combinations of hypothetical common traits or factors, y, that account for covariation among the variables and unique factors, e, that account for some variance of x but not covariance. This model can be written as

$$x = Ay + Be \qquad (6)$$

where x is a vector containing observed variables x_i, A is the n-by-k matrix of loadings on the k common factors, y is a vector containing k common factor scores, y_i, B is an n-by-n diagonal matrix with loadings along its diagonal and e is a vector containing the n unique factors e_i. The unique factors are assumed to be mutually uncorrelated and uncorrelated with the common factors. For item response data, the x factors can be thought of as unobservable response propensity variables that underlie the observed dichotomous responses. Let α_{ij} denote the loading of the i^{th} response propensity variable on the j^{th} factor and let β_i denote the single loading of the i^{th} response propensity variable on the i^{th} unique factor.

Lord and Novick (1968) show that for the two parameter normal ogive IRT model in unidimensional data,

$$a_i = \frac{\alpha_i}{\sqrt{1 - \alpha_i^2}}$$

or the item discrimination parameter for item i, a_i, is directly related to the factor loading α_i. They also show that

$$b_i = \frac{\gamma_i}{a_i}$$

or the item difficulty for item i, b_i, is directly related to the z transformation of the common item difficulty statistic (proportion of population that gives <u>incorrect</u> response). Drasgow and Parsons (1983) demonstrated that these relations hold very well in sample data generated for a three parameter logistic IRT. Therefore, common factors and factor scores can be used to create multidimensional simulation data.

In the proposed study, the common factor loading matrices, A, are specified to represent different characteristics of tests. First, factor loadings will range between .40 and .80 in the A matrix. This range was selected to represent items with low to high a parameters. Table 1 presents the loading matrix to be used in the present study. To achieve this loading matrix in aptitude test data, an oblique rotation would probably be necessary. Let Φ represent the matrix of first order factors after rotation to the simple structure in Table 1. It is fundamental to the simulation model presented here that there is one second order factor underlying the matrix of first order common factors.

Schmid and Leiman (1957) describe a hierarchical factor model that is based on a transformation of the k oblique first order factors to a matrix with k+1 orthogonal factors where the additional factor is the general factor represented in terms of item response propensity variables rather than first order factors. Basically, the larger the values in Φ, the stronger the general factor or the less the factor differentiation.

The hierarchical factor matrix with k+1 columns is used to compute item response propensity scores, x_i. To do this, a vector of k+2 factor scores is generated to represent the general factor, the k common first order factors, and the e_i (unique) factor. Factor scores are generated as independent, normal variables by the International Mathematical and Statistical Library (IMSL) Fortran subroutine GGNPM. The mean of the factor score distribution can be varied for different subpopulations. Factor scores are linearly combined according to eq. (6). The continuous variable, x_i, is transformed to a dichotomous variable, u_i, as follows. First, as part of the simulation model, each item, i, is assigned a value, denoted by γ_i, that is related to the desired proportion of examinees knowing the correct response (p). Then the continuous response propensity variable, x_i, is compared to γ_i and

$$u_i = 1 \text{ if } x_i \geq \gamma_i$$

and

$$u_i = 0 \text{ if } x_i < \gamma_i.$$

γ_i equals that point on the abscissa of the normal distribution (mean 0, standard deviation 1) to which the desired proportion of the population knowing the correct response (p) is represented by the area under the curve to the right of γ_i and 1-p lies to the left of γ_i. See Lord & Novick (1968, p. 370) for further discussion of this transformation. For the proposed study, the γ_i are presented in Table 1. These values were specified to represent items where the proportion of examinees in the majority population, knowing the correct response to an item, are .1, .2, .3, .4, .5, .6, .7, .8, .9. The values are equally represented among items by randomly selecting without replacement from 50 γ_i values and assigning them sequentially to the 50 items. For any desired factor model, a simulated examinee's item score vector, U, $(u_1, u_2, \ldots , u_n)$ can be obtained by generating a vector, y, computing $P(\theta)$ for each item and comparing this value to a random number drawn from a uniform distribution in the interval [0,1].

Item bias can be considered in data sets that earlier studies suggested had tolerable multidimensionality, assuming that more extreme cases would be identified as inappropriate by statistical procedures such as Reckase's

(1979) suggestion that the first principal component equal at least 20% of the total variance in a test or Drasgow and Lissak's (1983) parallel analysis method that requires more computation, but is more compelling than a 20% rule of thumb. The model used in the current study is based on a general factor (G) and two common factors (F_1 and F_2) where the population correlation matrix actually has rank 2 when communalities are in the diagonal of the matrix. The representation by three factors is the hierarchical factor structure that was developed by Schmid and Leiman (1957) and subsequently used by Humphreys (1962) in studies of human performance on cognitive ability tests.

For the current study, the two common factors can be thought of as verbal ability and quantitative ability underlying performance on a test of mathematical reasoning with word problems. In the population of individuals, the fairly high correlations between these two factors and therefore communality between them can be represented by the general factor with the respective unique components being orthogonal to each other and to the general factor. We are interested in measuring the general factor although the unique portions of verbal ability and quantitative ability affect performance on the test. In this context, items can be biased in a number of ways and will serve as the example for describing these ways.

Suppose, for instance, that the intercorrelation between common factors F(verbal) and F(quant) is very high. Then the unique variance attributed to each common factor is trivial and for item parameter purposes, the test and resulting item response data are almost unidimensional. Note, a correlation of 1.0 between factors reduces to a single factor.

On the other hand, if the factors are only moderately correlated, then the general factor will be smaller though still present (by definition of the model). Item parameter estimates for a unidimensional model in this multidimensional data will be subject to two types of error. One is conceptually based and derives from the fact that a single factor or dimension simply cannot capture completely the systematic effects on item responses. This will be termed lack-of-fit error. The second type of error is sampling error. The lack of complete specification of ability by the unidimensional model leads to less reliable estimates of all parameters in the model. Analogously, estimation of item parameters is more accurate if true ability is known. When estimates of ability must be part of the total parameter estimation procedure, then logically there will be more error in item parameter estimates.

In the above cases, we could consider two subpopulations and how an item could be biased against one subpopulation or the other. In one regard, the model is inappropriate for both groups (due to lack-of-fit), but not biased towards either group with regards to the other. However, larger parameter estimation errors that occur because of multidimensionality could lead to large observed ICC differences than under the condition of perfect model validity.

Next, consider the cases where the distribution of abilities underlying the factors differs between subpopulations. The unidimensional bias indices might identify appropriately multidimensional items as biased as follows. In the example where quantitative ability and verbal ability correlate rather highly, the common variance between factors is most closely associated with test performance and hence the construct, mathematical reasoning. However, if subpopulation differences in ability distributions are present for the unique portion of verbal ability, then the lower verbal ability group would have lower probabilities of success on the items, even though the general factor loadings and general ability distributions are identical.

Methods

Simulated item response data were generated in the same manner as the Drasgow and Parsons (1983) and Parsons (1982) studies. A factor population factor model was specified with two oblique factors. The degree of correlation between factors was varied. Since higher correlation implies less factor differentiation, a factor intercorrelation of .5 was set as high differentiation (but within tolerable limits for logistic parameter estimation). Low differentiation will refer to a factor intercorrelation of .81. Factor intercorrelation of .50 was the approximate lower limit found by Drasgow and Parsons (1983) for a meaningful estimation of θ in multidimensional data.

Along with the two levels of factor differentiation, a subpopulation difference on a minor factor (secondary ability) was either present or not. Note that IRT parameter invariance properties have been defined for a unidimensional θ which tends to represent the general factor in multidimensional data. Therefore, the equating of ability distributions between subpopulations does not directly affect known differences in secondary ability distributions. This model implies exactly the same factor structure matrix for both subpopulations, with only the distribution of secondary ability differing between groups. These population structure matrices appear in Table 1. The oblique factor matrix is constant throughout all four cases. Only the transformation matrix affects the size of the loadings on the general and orthogonal common factors.

Hypothetical ability vectors were created by using the IMSL subroutine GGNPM to generate four random factor scores from a normal distribution with mean=0 and standard deviation of 1. One random normal deviate represents an individual's general ability score, one deviate represented an individual's ability on factor 1 (F_1) score, one deviate represented an individual's ability on factor 2 (F_2) score, and one deviate represents the factor score unique to each item. If the vector of three common factor scores is represented by "y" and the factors that are unique to an item are represented by "e," then a vector of response tendency scores, x, was determined according to eq. 6.

To clarify the combinations of factor correlations and ability distributions to be analyzed, I will nominally label low factor differentiation and no secondary ability difference as Case 1, low factor differentiation and mean secondary ability difference of 1.0 as Case 2, high factor differentiation and no secondary ability difference as Case 3, and high factor differentiation and mean secondary ability difference of 1.0 as Case 4. To specify the true effect of these combinations on the probability of correct answer to each item by different subpopulation, the item response probability surface can be approximated by assuming that for any given ability level for the general factor, the secondary abilities will have a mean ability level equal to their mean. That is, the distributions are multivariate normal and variance is homoscedastic. The "average" item response surface is then reduced to a line and an item characteristic curve. For example, when G=-3.0, the mean F1 score is -.5. Therefore, the correct response propensity on item 1 in Table 1 is .021 when G=-3.0. If the mean F1 score is +.5 (a different subpopulation), then the probability of a correct response on item is .033 when G=-3.0. The differences between each pair of item response surfaces (different subpopulations) were computed between G=-3.0 and G=3.0 at intervals of .02, the difference multiplied by .02, and summed over the 301 intervals or:

$$\text{ALGDIF} = \sum_{i=1}^{301} \left[\left(P_{1i}(G) - P_{2i}(G)\right) \times .02\right] \qquad (8)$$

These differences can serve as population level item bias indices for the true underlying item response model. Note that when there is no difference in mean ability on either F1 or F2, there will not be any difference in the true item response surfaces. The population level bias indices ranged from .194 to .484 in Case 2 (low factor differentiation) and from .302 to .994 in Case 4 (high factor differentiation). In both cases, bias was present for the first 25 items only.

For each Case, two samples of 1,000 simulated item response vectors were generated for a total of eight samples. The sample size was chosen for the two parameter model because earlier studies by Drasgow and Parsons (1983) had demonstrated good parameter estimation with this sample size. For each case, the two samples were input to LOGIST, a maximum likelihood parameter estimation program (Wood, Wingersky, & Lord, 1976). Default program parameters were chosen in all cases except the choice of the "c" parameter which was fixed at zero.

All parameter estimates were equated between subpopulations before any comparisons were made or IRT bias indices computed. Since the scaling of Θ is arbitrary and LOGIST estimates of Θ have a mean 0 and standard deviation 1, item parameters were rescaled within each subpopulation so that the _b_'s had mean = 0 and standard deviation = 1. The equating constant for each sample were "SD" = Standard Deviation of _b_'s and "MN" = mean of _b_'s after eliminating all extreme _b_'s (absolute value greater than 3.0). Then

$$b^* = \frac{b - MN}{SD}$$

and

$$a^* = a \times SD$$

where b^* and a^* are the transformed values.

In past studies, the computation of IRT curve differences is as follows. First, because the _c_ parameter is frequently poorly estimated, Lord (1980) has recommended that samples be combined for estimation of _c_'s to obtain more stable estimates. Then, _a_'s and _b_'s are estimated in each sample with the _c_'s fixed at their earlier estimated level. This procedure suggests that bias in the _c_ parameter will not usually be detected because of sampling error. For this reason, in the current study, _c_'s were fixed at 0 for both generating item responses and estimating the parameters. This special two parameter case of the three parameter model does reduce the generalizability of the results, but should reduce the number of response vectors necessary for parameter estimation and also the cost of the parameter estimation.

The item bias statistics can then be computed by comparing equated estimated item characteristic curves. As in the true item response functions, the estimated curves were compared in the interval of $\Theta=-3$ to +3. The difference between subpopulations in estimated response probabilities were computed at intervals of .02 and summed across the 301 intervals. Both the algebraic sum and the sum of the absolute value of the differences were computed. The formulas were given in equation 2 and equation 3. These summed area differences were computed for the four cases described above.

In addition to the IRT item bias index, the method based on the proportion of individuals in each sample who got each item correct in five total score intervals was also used. This is the full Chi-Square method suggested by Camilli et al. (1980) which has received some empirical support as a reasonable approximation to the more costly IRT models. The index is given in equation 4 and equation 5.

Results

The results will be presented on a case by case basis with comparisons and generalizations made at the end.

Case 1 results concern the case where there is low differentiation and no mean difference on F_1. As noted in the discussion of true differences in item response surfaces, there is no population level bias in Case 1. Any observed large values in item bias statistics is solely due to lack of model fit and sampling error. Table 2 presents the estimated item parameters, the chi-square statistics, and the IRT curve differences, In general, the estimated parameters appear similar between the two samples. Figures 1 and 2 show the relationship between equated _a_ and _b_ parameter estimates respectively. Note that the estimated _a_ parameters show more scatter than the estimated _b_ parameters. The correlation between _a_'s is .93 and between _b_'s is .98. Thus, at the level of item parameter estimates, there seems to be a good deal of agreement between the two samples.

Averaging the four item bias indices for the first 25 items and then the second 25 items in Table 2 shows that there is little difference for any of the bias indices. For the signed chi-square index (ALGCHI), the mean is -.300 for the first 25 items and -.636 for the second 25 items. For the unsigned chi-square index (ABSCHI), the means are respectively 5.26 and 4.71. For the signed IRT index (ALGICC), the means are .013 and .032 respectively. Finally, the means for the unsigned IRT index (ABSICC) are .135 and .147. The distinction between first and second 25 items is only for comparison to Cases 2 and 4 where there was a difference in the response probability surface.

Case 2 results are for the situation where there is low factor differentiation and a mean difference in F_1. The item parameter estimates and bias indices appear in Table 3. The correlation between equated _a_'s is .92 and between equated _b_'s is .98. Figures 3 and 4 show the relationship between the estimated _a_'s and _b_'s respectively.

The signed IRT index and the signed chi-square index tend to be higher in the first 25 items (mean ALGCHI = 11.52, mean ALGICC = .086) than the second 25 items (mean ALGCHI = -9.81, mean ALGICC = -.166). The unsigned indices are both about equal in the two sets of items. For the unsigned chi-square index, the first 25 items had an average chi-square equal to 14.46 and the second 25 items had an average chi-square of 11.48. This difference is in the expected direction but both tend to be above the critical value for chi-square. For the unsigned IRT index, the mean was .174 for the first 25 items and .182 for the second 25 items. Based on these results, it appears that both the signed indices and possibly the unsigned chi-square index are effective when there is only slight differentiation between factors, but actual bias due to a subpopulation mean difference on a factor other than the general factor.

In case 3, there is a higher degree of factor differentiation and therefore poorer model fit. However, there is no population level bias for any of the items. Table 4 shows the item parameter estimates and bias statistics. The correlations between estimated a's is .92 and between b's is .98. Figures 5 and 6 show these plots and show that there are no outliers and generally better agreement for the b's.

When averaged for the two sets of 25 items, the bias statistics in Table 4 show that there is virtually no mean difference for ALGICC (.080 vs. .030) or for ABSICC (.149 vs. .137). For the Chi-Square indices, the mean differences were small for ALGCHI (.590 vs. -1.931) and somewhat larger for ABSCHI, but in the wrong direction (3.82 vs. 5.53). Again, there is no population level bias to explain this so it can be considered sampling error.

The final results are for Case 4 which had high factor differentiation and a mean difference of 1.0 in F_1. Table 5 presents the estimated item parameters, the two ICC difference statistics, and the two chi-square statistics for the 50 items. Note that in general, the estimated parameters again appear somewhat similar. The comparison of the transformed _a_'s in the two samples is shown in Figure 7. The correlation between _a_'s was .90. The comparison between _b_'s is shown in Figure 8. The correlation between _b_'s was .96.

The means for the two item sets on the four bias indices in Table 5 also show that both IRT indices and the signed chi-square index tend to be higher for those items that were inherently biased (the first 25 items) than for those items that were unbiased (the second 25 items). For ALGICC, the means were .326 and -.076. For ALGCHI, the means were 17.68 and -16.62. For ABSICC, the means were .340 and .187. On the other hand, for ABSCHI, the means were 18.97 and 17.29, which does not suggest much detection effectiveness.

Next, some analyses will be conducted to compare the results across cases. As mentioned in the introduction, item bias statistics are used to _identify_ items that are biased. Two types of errors can be made in this identification. First, failing to identify a biased item as such is false negative error and falsely identifying a non-biased item as biased is a false positive error. Detection at the chance level means that items are labelled as biased in proportion to their presence in the test. That is, if 10 items were biased in a 50 item test and a bias index only detected at the chance level, then for any number of items labelled as "biased", on the average .20 (10/50) would be actually biased and .80 (40/50) would not be biased. By rank ordering the items by each item bias statistic and noting the proportion of biased items identified as such at various levels of chance detection, we can gain some idea of the effectiveness of each. These analyses will only be conducted for all four cases even though only in Cases 2 and 4 was there actually any population level bias.

Figures 9, 10, 11, and 12 represent the results for the respective cases. Note first of all that in Figures 9 and 11, all four bias indices tend to follow the diagonal which represents chance detection. This is appropriate because there were no population biased items and the first 25 items were only nominally labelled as such. Therefore, it appears that none of the indices is over or under identifying the two sets of 25 items when there is no inherent bias.

Looking at Figure 10, it is clear that the two signed indices (ALGCHI and ALGICC) lie above the chance diagonal with neither one clearly superior to the other. The unsigned chi-square index, ABSCHI, does better than chance at low numbers of items identified, but not the higher numbers. In Figure 12, the two IRT indices and Chi-Square signed index again rise above the diagonal, showing good detection. The unsigned IRT index is slightly below, but still better than chance. Only the unsigned chi-square index appears to detect only at chance levels.

There is another perspective to consider on the various bias indices in the four cases. If an investigator has no a priori knowledge of the group that is disadvantaged by the items on a test, then the final sign associated with a computed index cannot be used to aid identification. For instance, if an item has a signed Chi-Square of 15.00 or -15.00, it will still be labelled as biased against one group or the other. In the cases where there was an ability difference (Cases 2 and 4), both signed indices tend to have larger negative values for items that are population unbiased. For chi-square, the mean value is -12.72 across Cases 2 and 4 (ability difference) as compared to -1.28 in Cases 1 and 3 (no ability difference). For ICC, the mean value is -.121 for an ability difference and .031 for no ability difference. For both chi-square and ICC, these values are large enough to cause to misidentification.

Discussion

This study examined the effects of multidimensionality and subpopulation differences in secondary abilities on commonly used item bias indices. The results were quite clear and suggest that in cases where the general factor is quite strong (though the test is not unidimensional) and where the general factor is not as strong (the test is clearly multidimensional, though all common factors are correlated), both the difference between estimated item characteristic curves and Camilli's (1980) full chi-square method will identify items biased because of subgroup differences on secondary ability. This study also suggests that the signed index is best in both cases. Both conclusions assume that the investigator has a priori knowledge of which group is at a disadvantage on the items.

The more traditional definition of item bias assumes a single latent ability IRT model where different subpopulations have different item characteristic curves. The current findings suggest that the unidimensional models can approximate multidimensional data and identify bias that is occurring in distributions of secondary abilities rather than item parameter differences. This is an important finding, because in real data, the unidimensional model is unlikely to be totally valid. In addition, if the general factor represents the only valid variance in observed test scores then bias that results from differences on secondary abilites is still bias and presents a disadvantage to the members of the group that is lower on this ability. To repeat, the notion of bias only holds if the secondary ability is irrelevant to the construct or criterion of interest.

As mentioned in the introduction, the three parameter logistic model has received most of the attention as a model for detecting item bias. The current study can be viewed as a special case where the _c_ parameter is held constant at 0. Drasgow and Parsons (1983) had studied the effects of multidimensionality on IRT parameter estimates and found that the two parameter model fit slightly better than the three parameter model. Therefore, the current effects would likely be similar for item bias indices in the three parameter model. As noted in the introduction, the recommended procedures for using the three parameter bias index includes setting the estimated _c_'s to a common value for the two groups suggesting that sampling error will swamp any true difference in _c_ parameters. Therefore, little or no generalizeability is lost by foregoing the estimation of the parameter in the current study. However, it is still an empirical question and worth future consideration.

Besides generalizing to the three parameter model, there are also many other issues worth studying in the general area of item bias and multidimensionality. For instance, it would be useful to know how the ratio of biased to unbiased items affects detection effectiveness. The current study used 25 items of each. A smaller number of item biased by subpopulation differences on a secondary factor would mean that the secondary factor itself was smaller. Therefore, the unidimensional IRT model would fit the data better. But the current results suggests that under the null condition of no biased items, simply increasing differentiation among factors does not appreciably increase bias indices. That is, lack-of-fit error by itself does not lead to misidentification of unbiased items as biased.

Probably the most important issue to be studied is how to interpret the signed indices when the investigator does not _a priori_ designate one group as disadvantaged. In multidimensional data, this would lead to correctly identifying some biased items, but also misclassifying some unbiased items. Earlier studies have not confronted this issue either.

In summary, the current study demonstrated that item bias resulting from subpopulation differences on a secondary ability will be detected as such. A non-trivial degree of multidimensionality can be tolerated by both the IRT and Chi-Square indices. Research on the interpretation of signed bias indices is still required.

References

Arvey, R. D. (1979). Fairness in Selecting Employees, Reading, Mass.: Addison-Wesley.

Berk, R. A. (Ed.) (1982). Handbook of Methods for Decting Test Bias. Baltimore, Maryland: The Johns Hopkins University Press.

Birnbaum, A. (1968). In F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Drasgow, F. & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response models to multidimensional data. Applied Psychological Measurement.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item Response Theory: Application to Psychological Measurement, Homewood, Ill.: Dow-Jones.

Ironson, G. H. & Subkoviak, M. S. (1979). A comparison of several methods of assessing bias. Journal of Educational Measurement, 16, 209-225.

Linn, R., Levine, M. J., Hastings, N. C., & Wardrop, J. (1981). Item bias on a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems, Hillsdale, N. J.: Lawrence Erlbaum Associates.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

Parsons, C. K. (1982). The robustness of the unidimensional item response model. Project Report for 1982 Air Force Office of Scientific Research (Summer, 1982). Conducted by Southeastern Center for Electrical Engineering Education.

Rasch, G. (1960). Probablistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monté Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Schmid, J. & Lieman, J. (1957). The application of hierarchical factor solutions. Psychometrika, 1957, 53-61.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Thurstone, L. L. (1947). Multiple Factor Analysis, Chicago: University of Chicago Press.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST-a computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6, Princeton, N.J.: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Table 1

Orthogonal and Oblique Population Structure Matrices

			Orthogonal Factors			Orthogonal Factors		
Item	Oblique Factor		High Differentiation			Low Differentiation		
	F1	F2	G	F1	F2	G	F1	F2
1	0.400	0.000	0.283	0.283	0.000	0.360	0.174	0.000
2	0.500	0.000	0.354	0.354	0.000	0.450	0.218	0.000
3	0.600	0.000	0.424	0.424	0.000	0.540	0.262	0.000
4	0.700	0.000	0.495	0.495	0.000	0.630	0.305	0.000
5	0.800	0.000	0.566	0.566	0.000	0.720	0.349	0.000
6	0.400	0.000	0.283	0.283	0.000	0.360	0.174	0.000
7	0.500	0.000	0.354	0.354	0.000	0.450	0.218	0.000
8	0.600	0.000	0.424	0.424	0.000	0.540	0.262	0.000
9	0.700	0.000	0.495	0.495	0.000	0.630	0.305	0.000
10	0.800	0.000	0.566	0.566	0.000	0.720	0.349	0.000
11	0.400	0.000	0.283	0.283	0.000	0.360	0.174	0.000
12	0.500	0.000	0.354	0.354	0.000	0.450	0.218	0.000
13	0.600	0.000	0.424	0.424	0.000	0.540	0.262	0.000
14	0.700	0.000	0.495	0.495	0.000	0.630	0.305	0.000
15	0.800	0.000	0.566	0.566	0.000	0.720	0.349	0.000
16	0.400	0.000	0.283	0.283	0.000	0.360	0.174	0.000
17	0.500	0.000	0.354	0.354	0.000	0.450	0.218	0.000
18	0.600	0.000	0.424	0.424	0.000	0.540	0.262	0.000
19	0.700	0.000	0.495	0.495	0.000	0.630	0.305	0.000
20	0.800	0.000	0.566	0.566	0.000	0.720	0.349	0.000
21	0.400	0.000	0.283	0.283	0.000	0.360	0.174	0.000
22	0.500	0.000	0.354	0.354	0.000	0.450	0.218	0.000
23	0.600	0.000	0.424	0.424	0.000	0.540	0.262	0.000
24	0.700	0.000	0.495	0.495	0.000	0.630	0.305	0.000
25	0.800	0.000	0.566	0.566	0.000	0.720	0.349	0.000
26	0.000	0.400	0.283	0.000	0.283	0.360	0.000	0.174
27	0.000	0.500	0.354	0.000	0.354	0.450	0.000	0.218
28	0.000	0.600	0.424	0.000	0.424	0.540	0.000	0.262
29	0.000	0.700	0.495	0.000	0.495	0.630	0.000	0.305
30	0.000	0.800	0.566	0.000	0.566	0.720	0.000	0.349
31	0.000	0.400	0.283	0.000	0.283	0.360	0.000	0.174
32	0.000	0.500	0.354	0.000	0.354	0.450	0.000	0.218
33	0.000	0.600	0.424	0.000	0.424	0.540	0.000	0.262
34	0.000	0.700	0.495	0.000	0.495	0.630	0.000	0.305
35	0.000	0.800	0.566	0.000	0.566	0.720	0.000	0.349
36	0.000	0.400	0.283	0.000	0.283	0.360	0.000	0.174
37	0.000	0.500	0.354	0.000	0.354	0.450	0.000	0.218
38	0.000	0.600	0.424	0.000	0.424	0.540	0.000	0.262
39	0.000	0.700	0.495	0.000	0.495	0.630	0.000	0.305
40	0.000	0.800	0.566	0.000	0.566	0.720	0.000	0.349
41	0.000	0.400	0.283	0.000	0.283	0.360	0.000	0.174
42	0.000	0.500	0.354	0.000	0.354	0.450	0.000	0.218
43	0.000	0.600	0.424	0.000	0.424	0.540	0.000	0.262
44	0.000	0.700	0.495	0.000	0.495	0.630	0.000	0.305
45	0.000	0.800	0.566	0.000	0.566	0.720	0.000	0.349
46	0.000	0.400	0.283	0.000	0.283	0.360	0.000	0.174
47	0.000	0.500	0.354	0.000	0.354	0.450	0.000	0.218
48	0.000	0.600	0.424	0.000	0.424	0.540	0.000	0.262

Table 2

Estimated Item Parameters and Item Bias Indices for Case 1

Item	Low		High		ALGCHI	ABSCHI	ALGICC	ABSICC
	a	b	a	b				
1	0.376	2.299	0.315	2.977	-4.048	5.452	-0.190	0.201
2	0.573	-1.236	0.494	-1.275	-3.803	5.677	-0.036	0.086
3	0.812	-1.426	0.724	-1.440	-4.703	6.963	-0.052	0.061
4	1.067	-0.421	0.932	-0.372	-4.358	5.334	-0.057	0.067
5	1.347	0.305	1.280	0.369	-1.289	1.926	-0.041	0.041
6	0.545	-2.817	0.425	-3.466	-3.666	6.363	0.167	0.204
7	0.538	1.090	0.507	1.087	2.787	4.709	0.029	0.029
8	0.739	-1.376	0.794	-1.409	7.715	11.004	-0.032	0.088
9	0.902	-0.134	1.001	0.001	-6.997	7.145	-0.102	0.125
10	1.255	1.712	1.514	1.655	-0.503	4.484	0.085	0.119
11	0.516	3.175	0.337	4.260	2.802	9.961	-0.136	0.243
12	0.725	-2.222	0.400	-3.544	-3.940	12.294	0.479	0.579
13	0.643	0.970	0.702	0.883	1.337	2.353	0.071	0.121
14	0.944	-0.734	1.142	-0.676	3.587	4.891	-0.073	0.141
15	1.147	-0.005	1.145	-0.034	2.783	3.008	0.008	0.025
16	0.470	-1.841	0.388	-2.228	-0.797	4.618	0.112	0.160
17	0.703	-2.266	0.584	-2.787	2.986	3.117	0.206	0.213
18	0.744	1.457	0.643	1.633	-0.159	0.786	-0.062	0.085
19	1.079	0.077	0.932	-0.095	10.771	10.940	0.104	0.108
20	1.249	-0.305	1.408	-0.324	2.672	3.318	-0.008	0.071
21	0.467	-3.276	0.368	-3.714	-2.494	2.494	0.011	0.122
22	0.534	-2.873	0.548	-2.726	-0.821	3.282	-0.154	0.156
23	0.691	0.840	0.751	0.940	-4.458	4.458	-0.050	0.103
24	0.836	1.311	1.020	1.260	-3.174	3.676	0.063	0.152
25	1.169	0.002	1.299	0.012	0.300	3.286	-0.017	0.071
26	0.329	-1.876	0.444	-1.379	0.498	10.302	-0.249	0.397
27	0.595	2.672	0.761	2.315	-6.403	10.466	0.216	0.268
28	0.899	-1.356	0.689	-1.649	0.693	3.066	0.124	0.177
29	1.041	-0.318	0.845	-0.400	2.968	5.869	0.031	0.093
30	1.236	0.366	1.329	0.278	3.790	4.148	0.060	0.074
31	0.366	-3.712	0.523	-2.825	1.711	1.985	-0.336	0.421
32	0.433	-0.075	0.439	0.104	-3.106	3.199	-0.122	0.126
33	0.682	1.550	0.762	1.429	0.270	1.818	0.107	0.145
34	0.881	0.899	1.041	0.750	3.946	4.817	0.115	0.158
35	1.635	-1.578	1.559	-1.506	-1.956	3.883	-0.112	0.112
36	0.466	-2.097	0.543	-1.705	-5.605	9.614	-0.268	0.298
37	0.550	0.614	0.505	0.569	1.601	1.676	0.042	0.046
38	0.709	-0.397	0.765	-0.470	2.556	2.556	0.024	0.093
39	0.929	-0.905	0.871	-0.806	-3.703	3.817	-0.106	0.106
40	1.459	1.592	1.249	1.763	-1.108	1.696	-0.065	0.071
41	0.455	1.121	0.422	1.320	-2.554	4.842	-0.084	0.084
42	0.598	2.598	0.638	2.385	1.048	2.208	0.161	0.168
43	0.723	-2.345	0.744	-2.306	-0.076	7.559	-0.104	0.109
44	1.186	-1.795	1.314	-1.699	1.605	6.739	-0.134	0.140
45	1.195	-0.376	1.388	-0.365	1.741	2.057	-0.030	0.089
46	0.374	-1.726	0.414	-1.337	-4.564	6.853	-0.246	0.264
47	0.533	1.837	0.672	1.615	-2.300	5.605	0.139	0.245
48	0.701	0.459	0.634	0.611	-6.268	10.019	-0.089	0.091
49	0.923	-1.255	1.039	-1.242	2.292	2.973	-0.061	0.107
50	1.313	-0.032	1.161	-0.010	-1.833	5.589	-0.025	0.038

Table 3

Estimated Item Parameters and Item Bias Indices for Case 2

Item	Low		High		ALGCHI	ABSCHI	ALGICC	ABSICC
	a	b	a	b				
1	0.458	2.215	0.380	2.017	-3.174	8.146	0.014	0.156
2	0.639	-0.770	0.561	-1.312	-5.741	8.506	0.157	0.175
3	0.710	-1.231	0.770	-1.732	-13.929	13.929	0.142	0.142
4	0.845	-0.148	0.902	-0.742	30.620	30.620	0.197	0.197
5	1.462	0.522	1.235	0.058	20.640	20.640	0.114	0.122
6	0.390	-3.519	0.424	-3.426	-1.850	8.668	-0.125	0.130
7	0.467	1.602	0.524	0.993	7.276	8.870	0.174	0.184
8	0.630	-1.368	0.656	-1.886	9.770	9.861	0.151	0.151
9	0.865	0.238	1.000	-0.220	7.410	9.310	0.111	0.124
10	1.382	2.051	1.243	1.408	26.633	26.638	0.221	0.221
11	0.429	3.630	0.382	3.444	2.319	2.542	0.019	0.071
12	0.614	-2.339	0.768	-2.296	1.910	3.914	-0.165	0.205
13	0.720	1.145	0.640	0.693	13.512	13.512	0.107	0.125
14	0.849	-0.559	1.010	-1.024	17.991	17.991	0.118	0.138
15	1.257	0.282	1.356	-0.251	28.944	28.944	0.158	0.158
16	0.465	-1.931	0.471	-2.086	-0.982	1.384	-0.063	0.063
17	0.618	-2.246	0.565	-2.869	3.029	3.029	0.188	0.190
18	0.664	1.725	0.724	1.207	4.993	5.009	0.137	0.140
19	0.824	0.304	1.049	-0.230	21.165	21.165	0.158	0.189
20	1.230	0.024	1.382	-0.597	44.414	44.414	0.213	0.213
21	0.371	-3.318	0.631	-2.660	6.535	6.535	-0.309	0.471
22	0.439	-3.300	0.538	-3.125	3.057	3.897	-0.153	0.195
23	0.665	1.285	0.696	0.723	9.209	9.209	0.169	0.169
24	0.971	1.571	1.177	0.932	18.721	19.246	0.219	0.223
25	1.010	0.259	1.240	-0.338	35.505	35.505	0.197	0.203
26	0.395	-1.508	0.419	-1.255	-12.430	12.691	-0.288	0.288
27	0.775	2.337	0.588	2.571	3.203	5.435	-0.260	0.294
28	0.791	-1.386	0.727	-1.456	-8.428	8.428	-0.127	0.129
29	0.832	-0.337	0.997	-0.429	-6.941	7.618	-0.114	0.140
30	1.239	0.321	1.220	0.243	-12.322	12.322	-0.125	0.125
31	0.549	-2.724	0.528	-3.007	-0.390	3.987	-0.002	0.029
32	0.469	-0.038	0.510	0.058	-15.494	15.494	-0.225	0.225
33	0.724	1.479	0.790	1.377	-7.035	7.036	-0.115	0.116
34	1.030	0.745	0.885	0.731	8.455	9.966	-0.165	0.172
35	1.238	-1.819	1.729	-1.584	-7.677	7.682	-0.314	0.317
36	0.374	-2.339	0.470	-1.831	-10.416	11.569	-0.366	0.387
37	0.526	0.641	0.527	0.494	-5.138	6.181	-0.080	0.080
38	0.694	-0.417	0.657	-0.522	-10.677	11.024	-0.106	0.107
39	1.044	-0.741	0.998	-0.818	-10.065	10.065	-0.123	0.123
40	1.254	1.665	1.799	1.475	-12.822	12.822	-0.060	0.130
41	0.323	1.602	0.385	1.423	-12.457	13.577	-0.107	0.171
42	0.555	2.991	0.473	3.133	1.740	3.859	-0.160	0.180
43	0.900	-1.965	0.874	-1.987	-5.678	5.678	-0.154	0.154
44	1.027	-1.813	1.214	-1.658	-6.261	6.696	-0.263	0.264
45	1.266	-0.383	1.507	-0.317	-36.489	36.822	-0.213	0.213
46	0.439	-1.277	0.403	-1.318	-6.015	6.015	-0.146	0.150
47	0.574	1.618	0.760	1.440	-16.932	16.932	-0.083	0.201
48	0.632	0.497	0.729	0.533	-22.701	22.701	-0.197	0.203
49	0.932	-1.285	1.002	-1.234	-9.080	9.161	-0.200	0.200
50	1.286	-0.057	1.398	-0.075	-23.223	23.223	-0.161	0.161

Table 4

Estimated Item Parameters and Item Bias Indices for Case 3

Item	Low		High		ALGCHI	ABSCHI	ALGICC	ABSICC
	a	b	a	b				
1	0.428	2.092	0.418	2.201	0.426	1.362	0.010	0.028
2	0.544	-1.138	0.486	-1.338	4.766	5.649	0.209	0.221
3	0.730	-1.423	0.697	-1.416	0.015	1.404	0.089	0.093
4	0.890	-0.382	0.996	-0.398	3.792	3.792	0.102	0.110
5	1.193	0.274	1.307	0.362	-1.045	1.528	0.022	0.041
6	0.452	-3.295	0.363	-3.782	-0.879	1.800	0.164	0.216
7	0.482	1.042	0.481	1.177	-6.447	1.049	-0.014	0.018
8	0.783	-1.562	0.696	-1.431	-2.069	6.701	0.000	0.094
9	0.902	-0.069	0.919	0.020	5.175	6.134	0.024	0.024
10	0.977	2.016	1.284	1.616	-0.968	6.064	0.338	0.343
11	0.368	3.801	0.443	3.384	-1.567	1.458	0.079	0.120
12	0.570	-2.658	0.368	-3.852	-1.508	7.111	0.473	0.536
13	0.624	0.935	0.694	0.892	1.560	3.881	0.098	0.114
14	0.831	-0.935	0.800	-0.780	-4.486	4.640	-0.016	0.037
15	1.119	0.009	1.206	-0.006	5.901	6.537	0.097	0.098
16	0.487	-1.898	0.434	-2.056	-0.291	1.125	0.152	0.177
17	0.585	-2.619	0.684	-2.453	3.884	3.886	0.024	0.098
18	0.594	1.558	0.718	1.312	0.708	4.051	0.210	0.233
19	0.843	-0.166	1.167	-0.055	-3.545	6.103	0.009	0.181
20	1.086	-0.367	1.111	-0.326	5.030	6.590	0.061	0.061
21	0.400	-3.615	0.531	-2.684	-1.157	1.848	-0.274	0.317
22	0.539	-2.707	0.692	-2.168	-0.606	1.758	-0.193	0.246
23	0.696	0.808	0.689	0.865	0.077	1.204	0.039	0.039
24	0.819	1.415	0.910	1.179	7.799	9.005	0.234	0.234
25	1.353	-0.046	1.314	0.006	0.191	0.763	0.050	0.051
26	0.333	-1.822	0.428	-1.272	-5.806	6.658	-0.177	0.277
27	0.402	3.443	0.606	2.521	-0.739	3.094	0.296	0.366
28	0.665	-1.613	0.560	-1.948	3.241	9.111	0.296	0.309
29	0.677	-0.598	0.719	-0.438	-1.477	7.116	-0.021	0.041
30	0.942	0.234	0.938	0.438	-5.093	5.093	-0.060	0.060
31	0.429	-3.685	0.330	-4.217	-3.378	4.137	0.101	0.191
32	0.417	-0.220	0.491	-0.006	-1.890	3.074	-0.063	0.156
33	0.548	1.637	0.622	1.543	-0.753	2.362	0.104	0.129
34	0.808	0.712	0.865	0.911	-6.034	6.034	-0.062	0.066
35	1.033	-1.935	1.031	-1.737	-7.158	9.308	-0.037	0.037
36	0.406	-2.269	0.463	-2.005	0.048	1.128	-0.028	0.107
37	0.412	0.571	0.562	0.674	-5.583	6.111	-0.032	0.277
38	0.565	-0.658	0.651	-0.486	-3.043	4.921	-0.027	0.114
39	0.889	-0.902	0.872	-0.861	0.877	1.945	0.066	0.066
40	0.856	1.935	1.089	1.840	-3.726	3.726	0.119	0.162
41	0.344	1.590	0.392	1.365	1.107	1.141	0.140	0.165
42	0.470	3.248	0.509	3.016	-0.906	4.308	0.113	0.115
43	0.853	-2.206	0.684	-2.291	-6.372	6.526	0.134	0.194
44	0.773	-2.421	0.685	-2.333	6.991	7.316	0.024	0.091
45	0.809	-0.564	0.932	-0.355	-2.375	10.799	-0.057	0.097
46	0.309	-1.867	0.337	-1.522	-1.696	3.083	-0.085	0.102
47	0.461	1.937	0.484	2.097	-1.589	3.031	-0.050	0.050
48	0.587	0.318	0.673	0.378	0.694	5.402	0.035	0.108
49	0.728	-1.527	0.799	-1.330	1.806	9.546	-0.036	0.065
50	0.913	-0.161	1.059	-0.016	-5.412	13.391	-0.015	0.080

Table 5

Estimated Item Parameters and Item Bias Indices for Case 4

Item	Low		High		ALGCHI	ABSCHI	ALGICC	ABSCHI
	a	b	a	b				
1	0.426	2.574	0.404	1.829	10.559	11.450	0.337	0.337
2	0.535	-0.767	0.530	-1.511	-13.158	10.421	0.315	0.315
3	0.616	-1.119	0.837	-1.697	23.376	23.376	0.222	0.301
4	0.850	0.003	0.813	-0.758	22.761	22.761	0.356	0.356
5	1.135	0.658	1.158	-0.023	22.457	22.457	0.320	0.320
6	0.421	-3.231	0.478	-3.310	1.240	4.849	-0.046	0.117
7	0.504	1.497	0.478	0.651	15.123	15.123	0.419	0.419
8	0.558	-1.373	0.692	-1.897	15.829	16.073	0.186	0.241
9	0.821	0.361	0.897	-0.368	20.310	20.310	0.344	0.344
10	1.255	2.043	1.195	1.366	11.764	11.768	0.349	0.349
11	0.488	3.393	0.453	2.827	4.570	5.764	0.243	0.243
12	0.528	-2.321	0.575	-2.929	8.271	9.191	0.208	0.208
13	0.596	1.455	0.675	0.591	15.110	15.110	0.433	0.434
14	0.890	-0.422	0.891	-1.215	26.078	26.063	0.368	0.368
15	0.964	0.399	1.178	-0.366	36.834	36.834	0.368	0.371
16	0.367	-2.100	0.453	-2.604	8.657	8.657	0.208	0.250
17	0.610	-2.182	0.634	-2.774	4.000	4.000	0.192	0.192
18	0.701	1.771	0.787	0.984	10.943	11.730	0.397	0.397
19	0.794	0.303	0.938	-0.390	26.688	27.380	0.318	0.322
20	1.003	0.045	1.076	-0.739	38.303	38.303	0.372	0.372
21	0.396	-3.191	0.272	-5.900	10.108	10.191	0.556	0.561
22	0.559	-2.263	0.609	-3.204	17.224	17.225	0.398	0.398
23	0.684	1.361	0.716	0.472	24.310	24.307	0.461	0.461
24	0.955	1.656	0.895	0.801	38.676	38.776	0.455	0.455
25	1.505	0.413	1.250	-0.340	42.037	42.037	0.361	0.361
26	0.349	-1.671	0.374	-1.565	-13.020	13.020	-0.179	0.190
27	0.570	2.724	0.502	2.887	-4.526	5.730	-0.115	0.123
28	0.715	-1.412	0.587	-1.740	-7.704	7.704	0.025	0.124
29	0.902	-0.364	0.693	-0.582	-16.029	16.237	-0.010	0.153
30	1.167	0.348	0.971	0.360	-42.443	42.444	-0.141	0.149
31	0.376	-3.916	0.357	-4.238	-3.146	5.248	-0.025	0.025
32	0.498	0.090	0.501	-0.035	-14.380	17.655	-0.056	0.062
33	0.650	1.434	0.736	1.457	-27.328	27.328	-0.132	0.155
34	0.883	0.702	0.731	0.907	-41.155	41.155	-0.255	0.259
35	0.909	-2.065	0.993	-1.880	-24.808	24.808	-0.304	0.305
36	0.495	-1.775	0.352	-2.829	0.989	5.493	0.313	0.384
37	0.503	0.408	0.397	0.651	-21.845	21.845	-0.251	0.290
38	0.527	-0.657	0.530	-0.646	-24.102	24.145	-0.157	0.157
39	0.837	-0.783	0.739	-0.920	-18.781	18.781	-0.071	0.086
40	0.973	1.951	1.124	1.824	-15.673	15.673	-0.018	0.090
41	0.283	1.833	0.440	1.301	-17.019	17.342	0.030	0.436
42	0.486	3.015	0.697	2.190	-7.502	7.518	0.283	0.366
43	0.587	-2.746	0.719	-2.520	-3.603	4.116	-0.275	0.298
44	0.909	-1.904	0.898	-1.980	-11.204	13.498	-0.134	0.134
45	1.063	-0.373	0.925	-0.385	-38.185	38.185	-0.143	0.146
46	0.391	-1.405	0.299	-1.792	-9.118	9.118	-0.007	0.220
47	0.505	1.994	0.538	1.853	-13.555	13.618	-0.026	0.076
48	0.829	0.404	0.588	0.395	-16.430	16.430	-0.119	0.249
49	0.808	-1.388	0.837	-1.460	-12.906	13.186	-0.124	0.125
50	1.076	0.078	0.909	-0.112	-12.018	12.018	-0.016	0.074

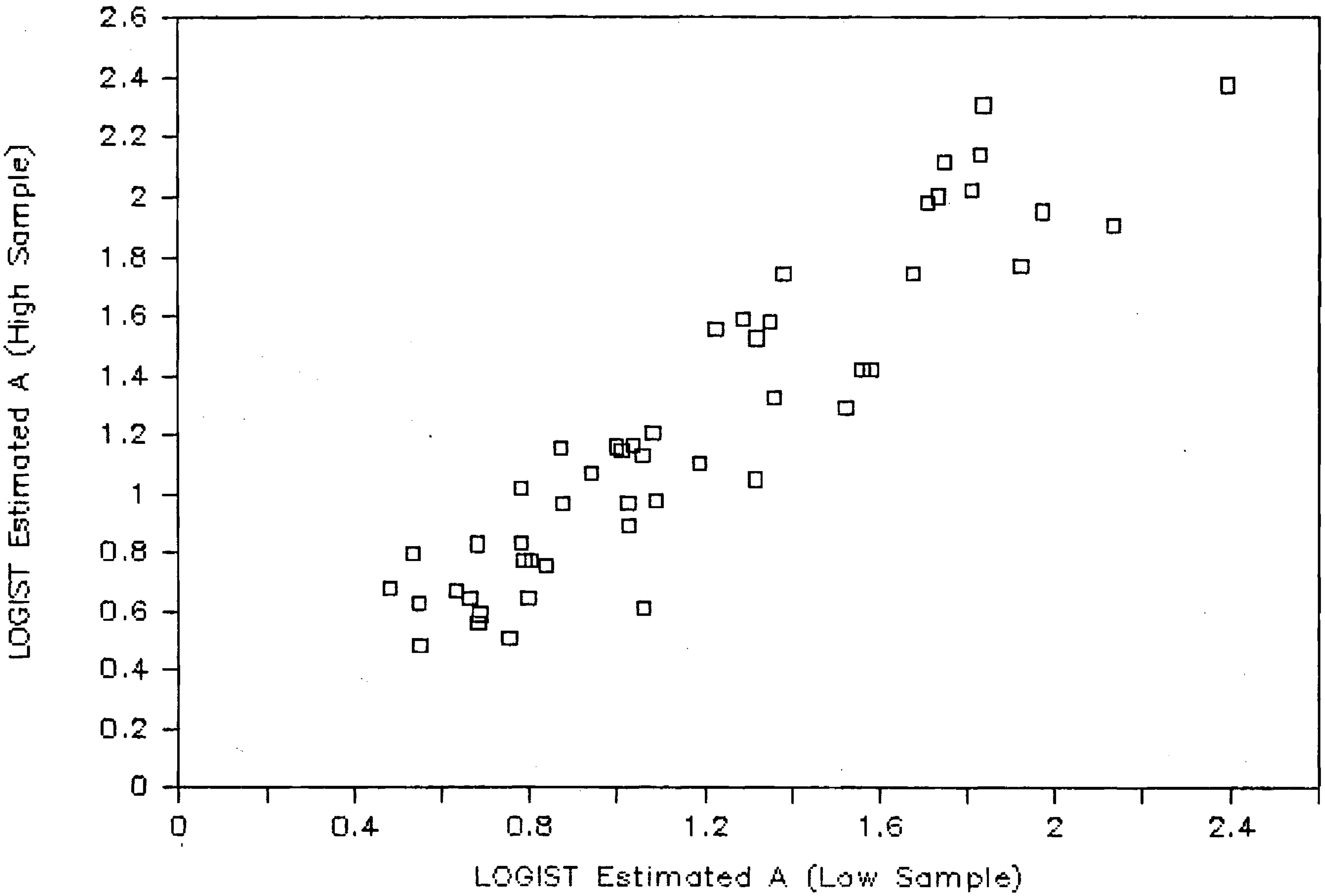

Figure 1. Comparison of Equated a Estimates for Case 1 Samples

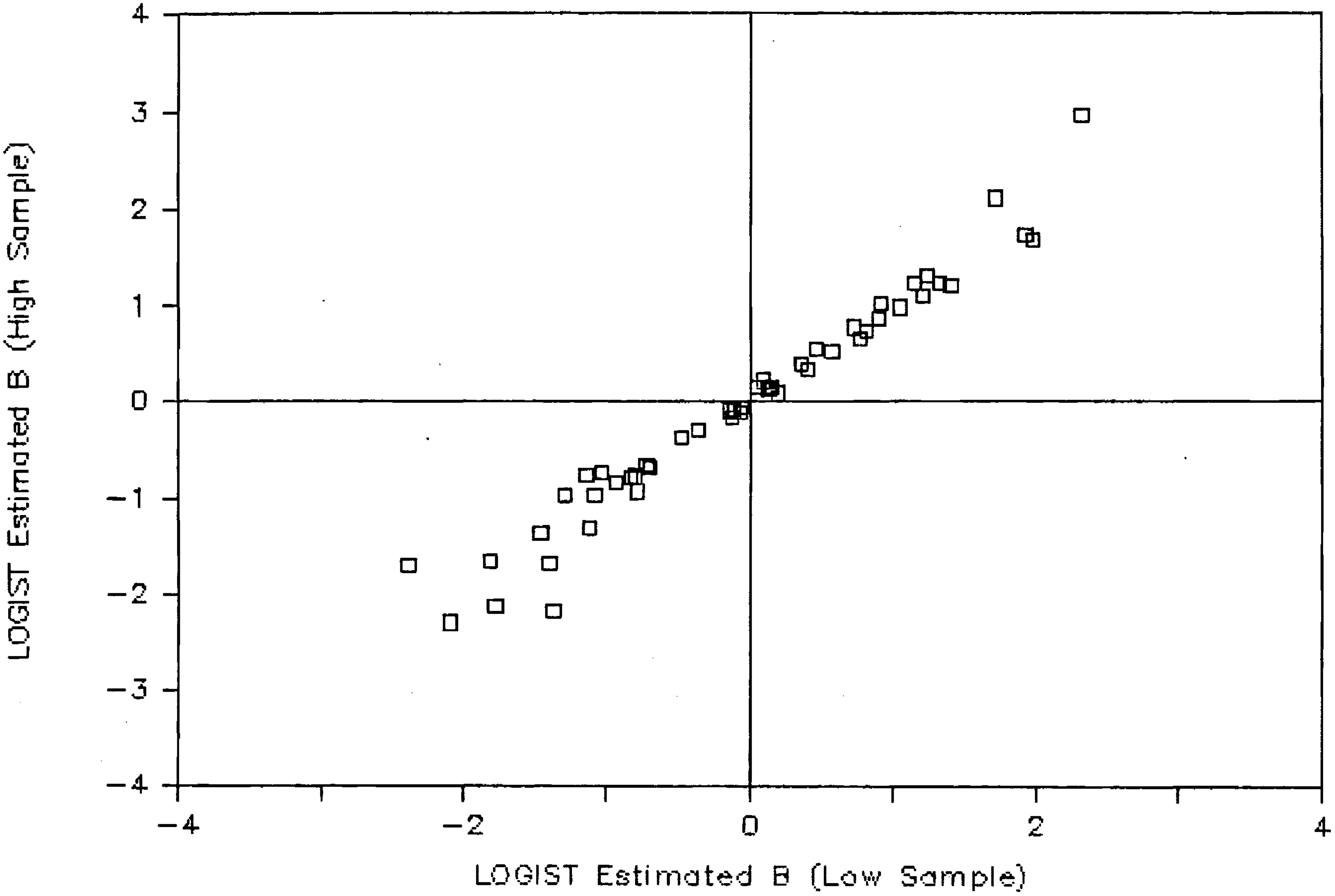

Figure 2. Comparison of Equated b Estimates for Case 1 Samples

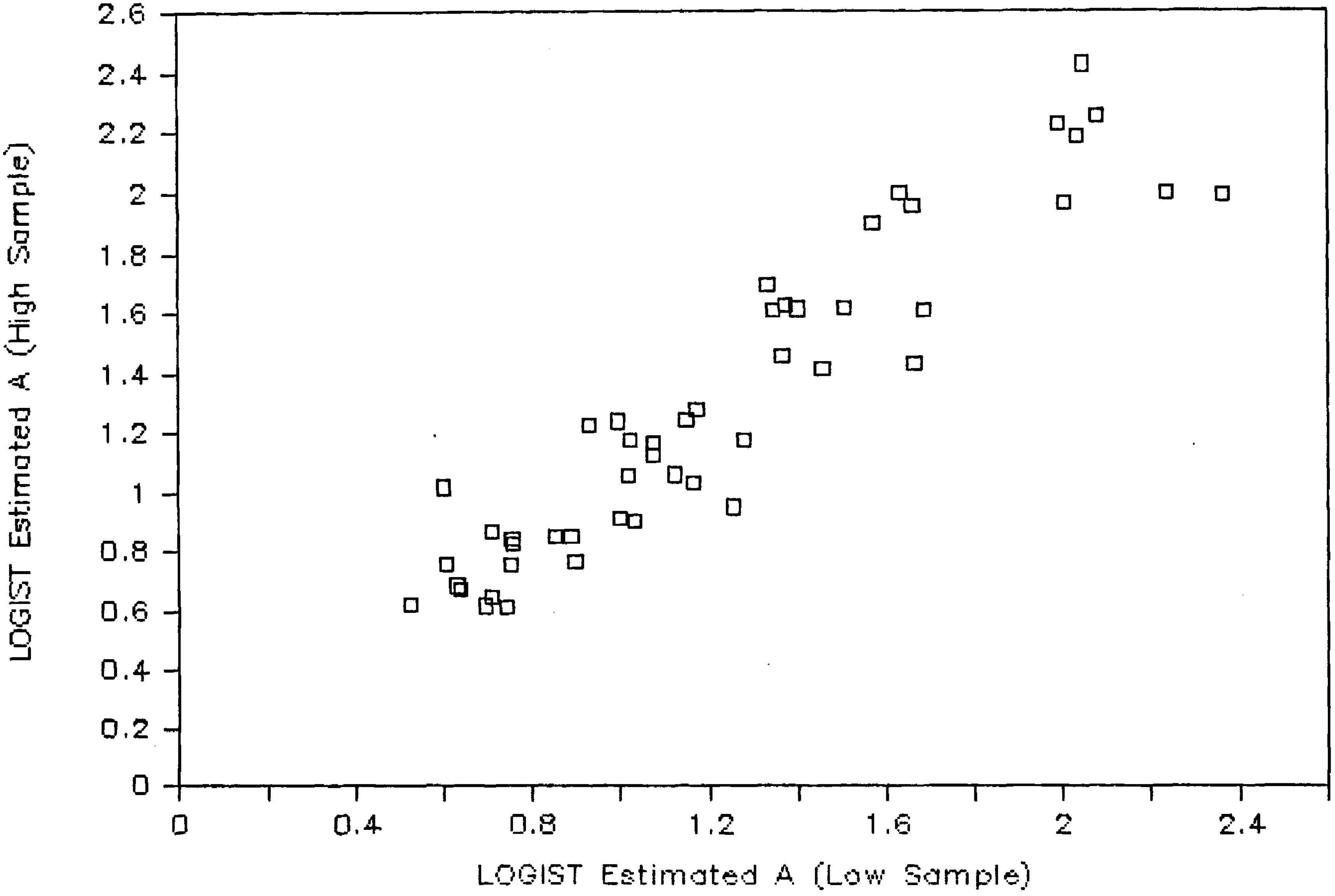

Figure 3. Comparison of Equated a Estimates for Case 2 Samples

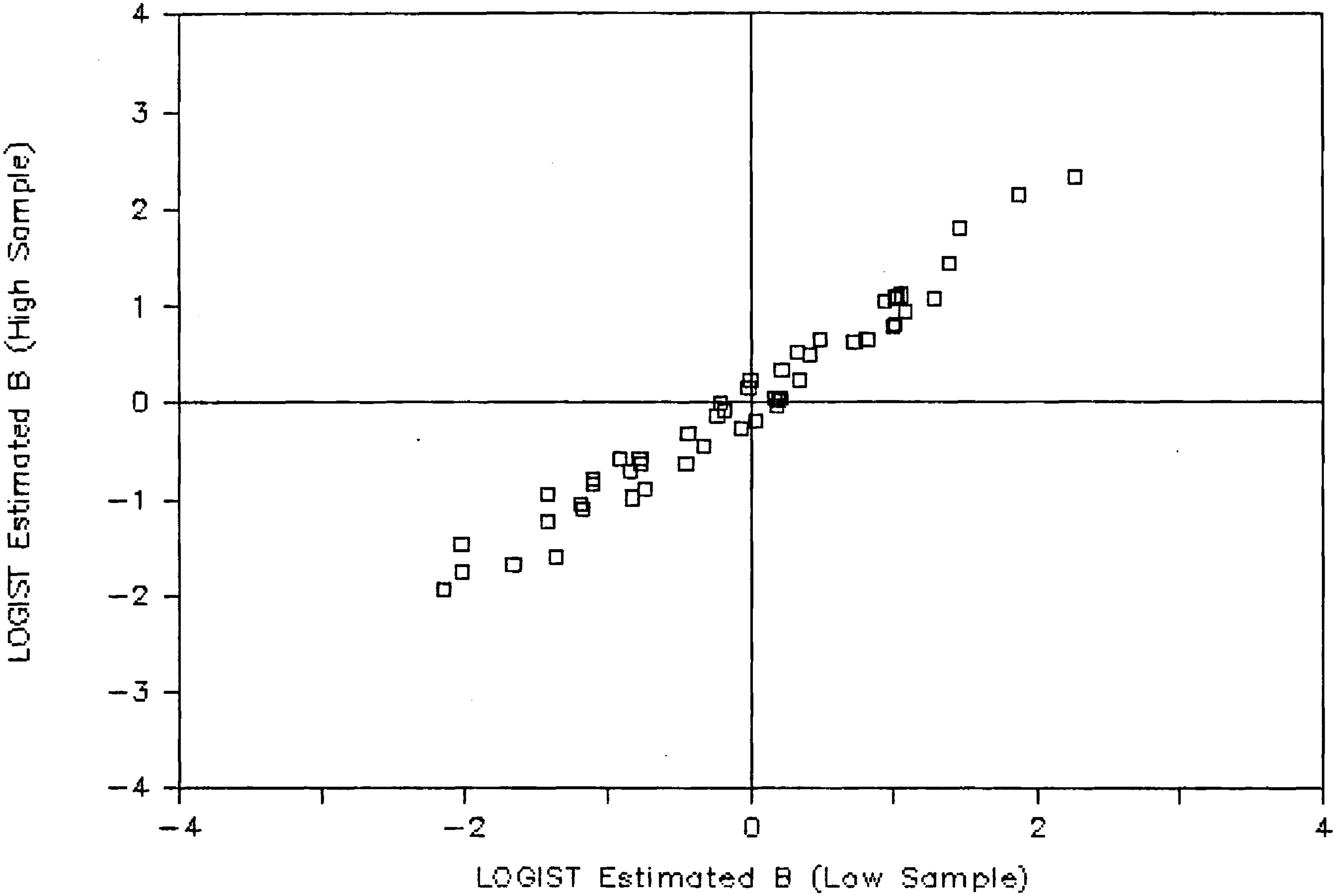

Figure 4. Comparison of Equated b Estimates for Case 2 Samples

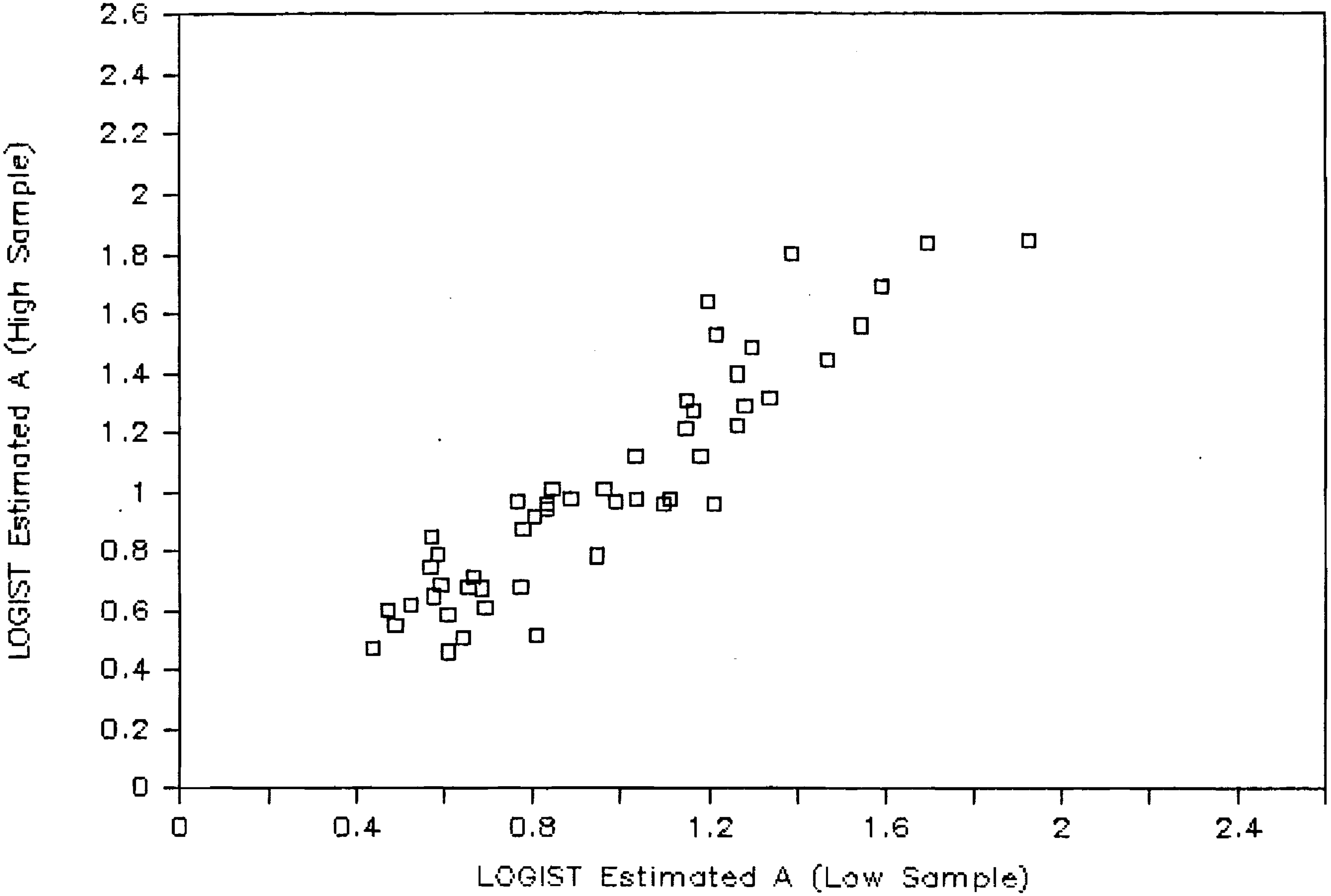

Figure 5. Comparison of Equated a Estimates for Case 3 Samples

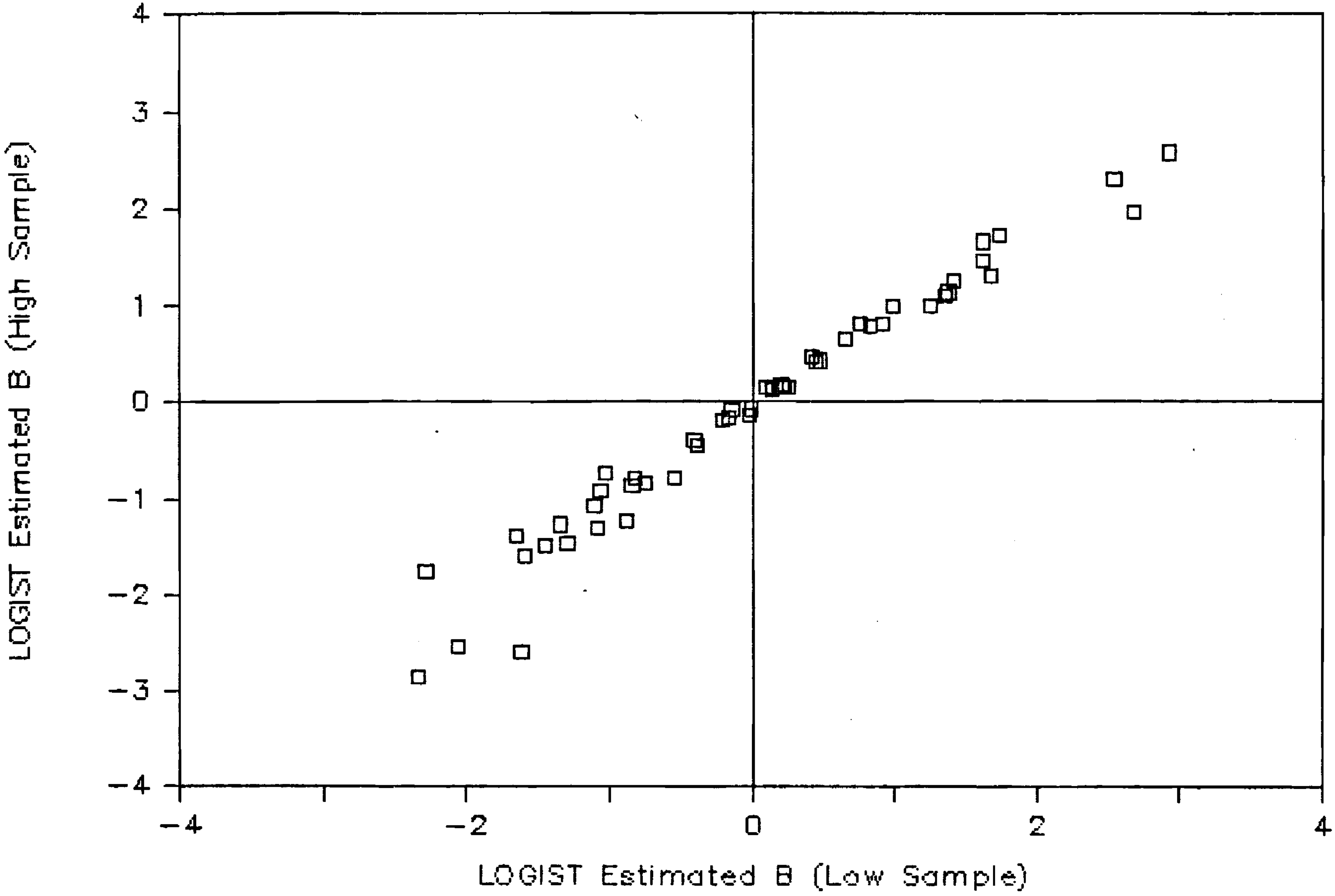

Figure 6. Comparison of Equated b Estimates for Case 3 Samples

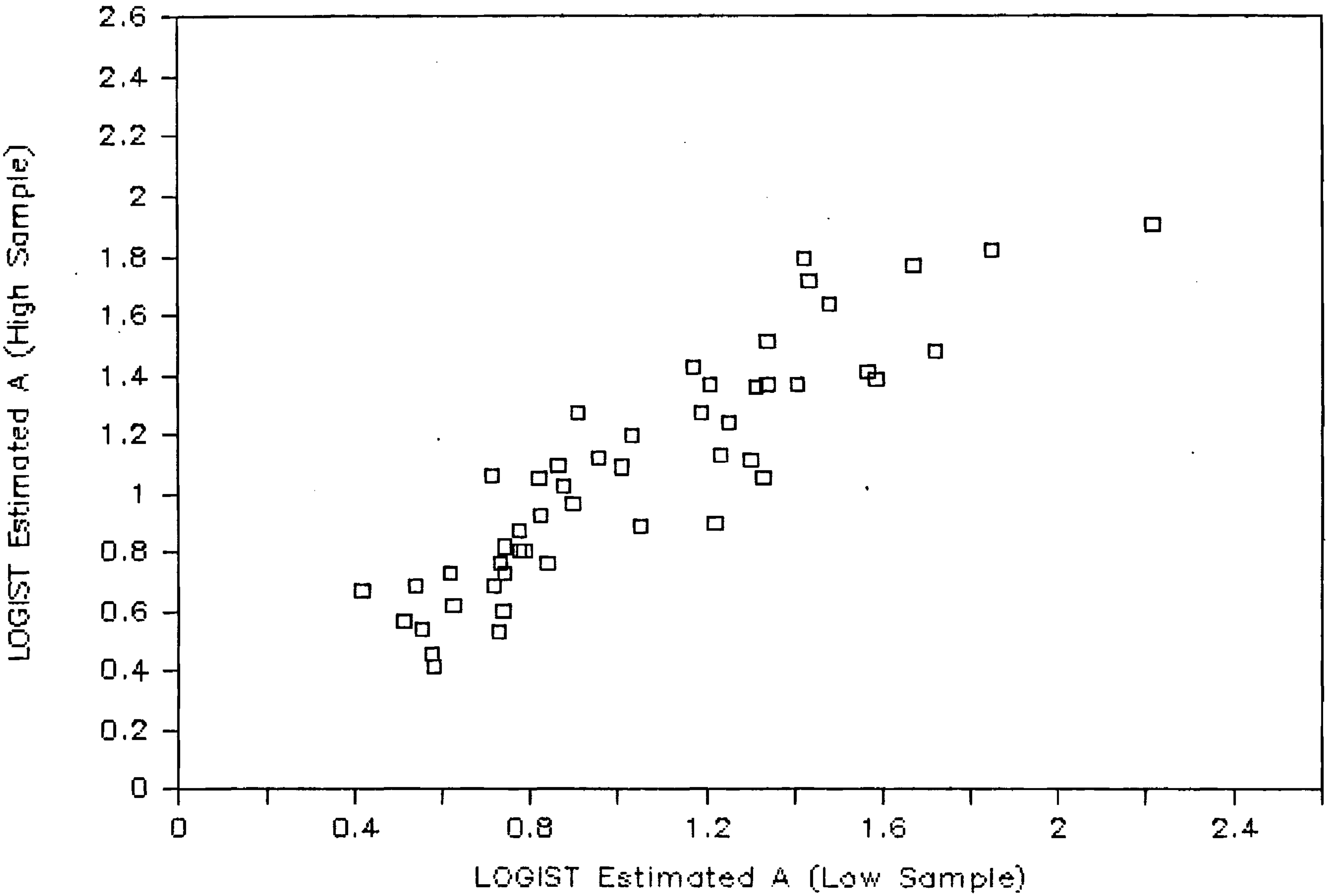

Figure 7. Comparison of Equated _a_ Estimates for Case 4 Samples

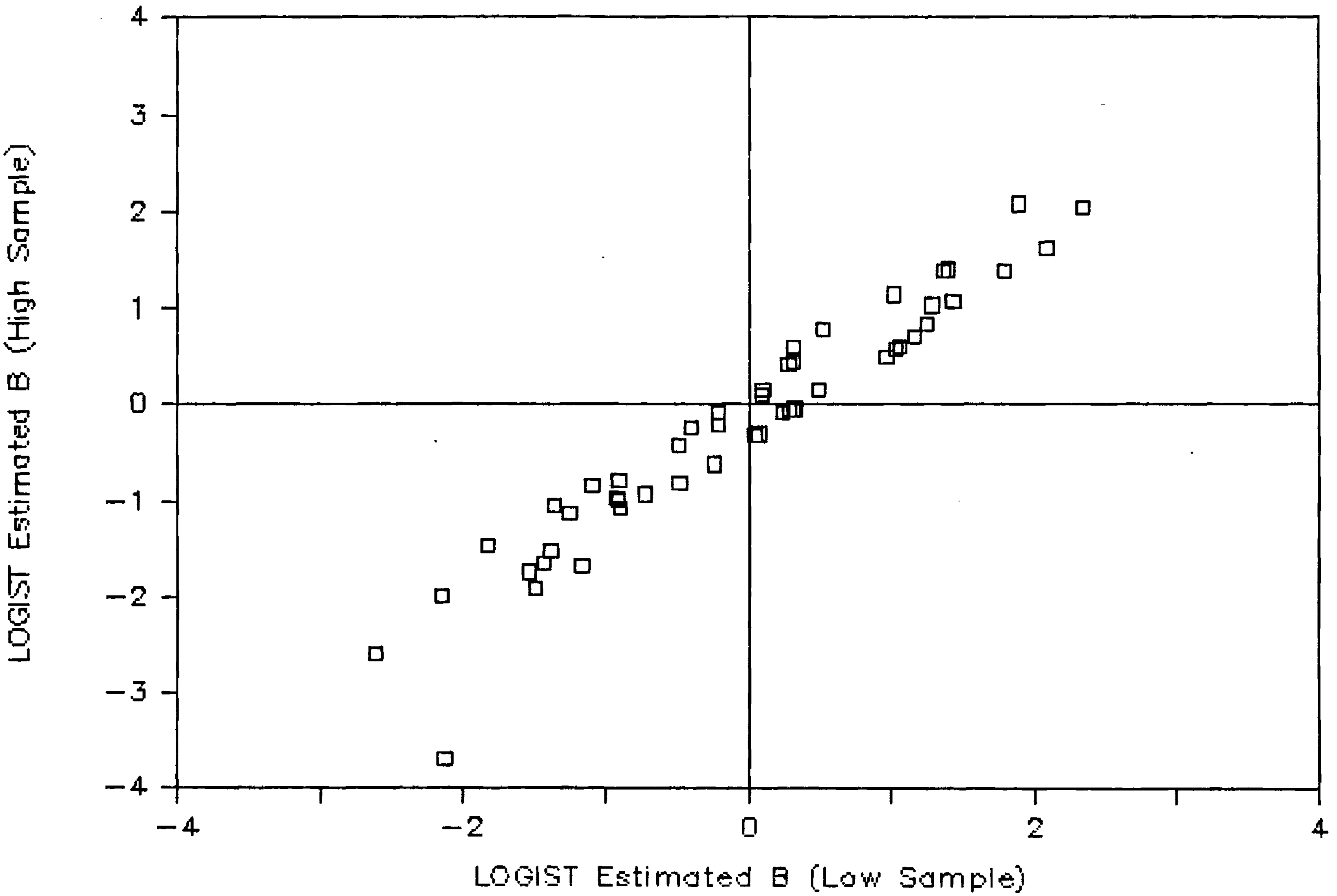

Figure 8. Comparison of Equated _b_ Estimates for Case 4 Samples

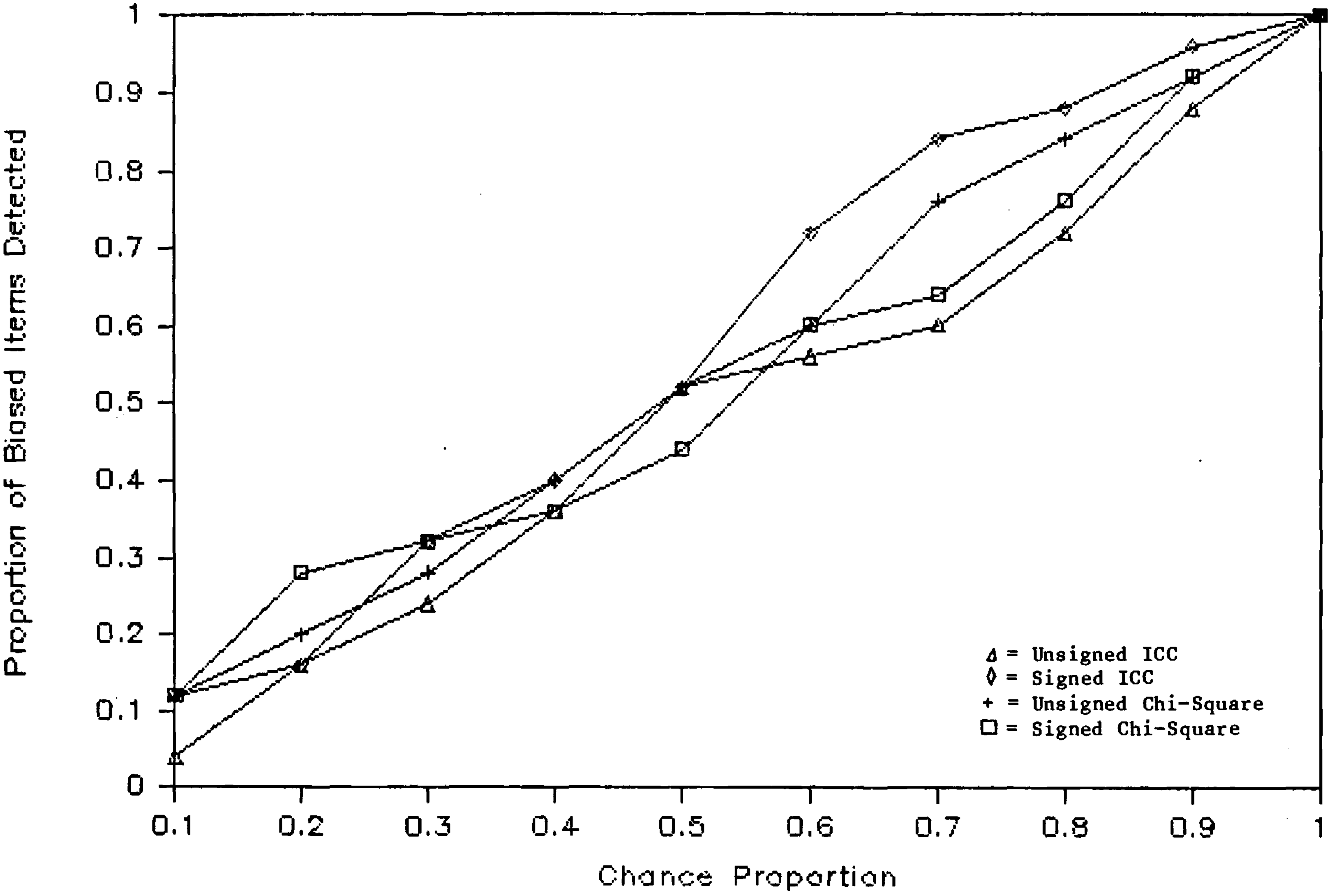

Figure 9. Bias Detection vs. Chance for Case 1

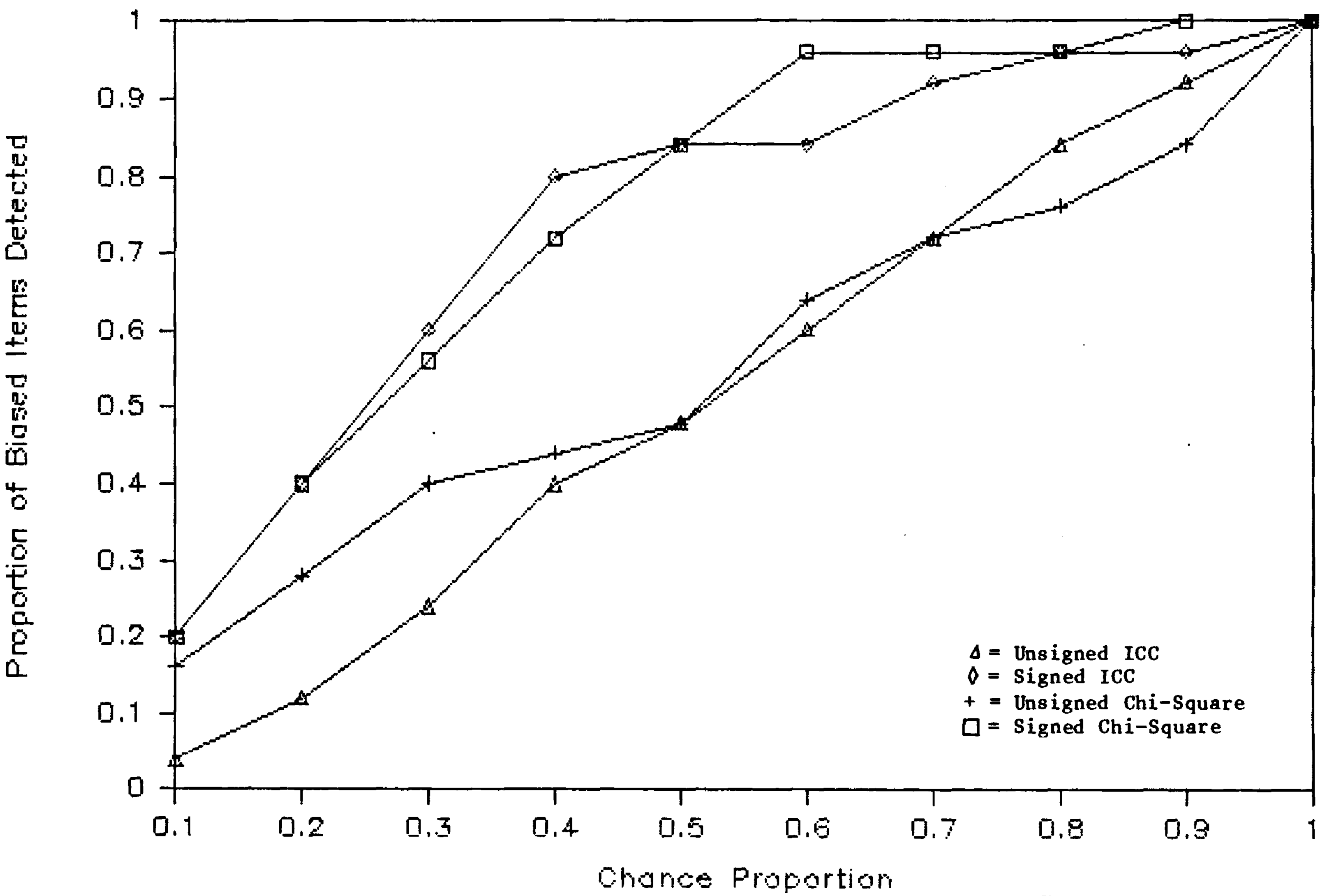

Figure 10. Bias Detection vs. Chance for Case 2

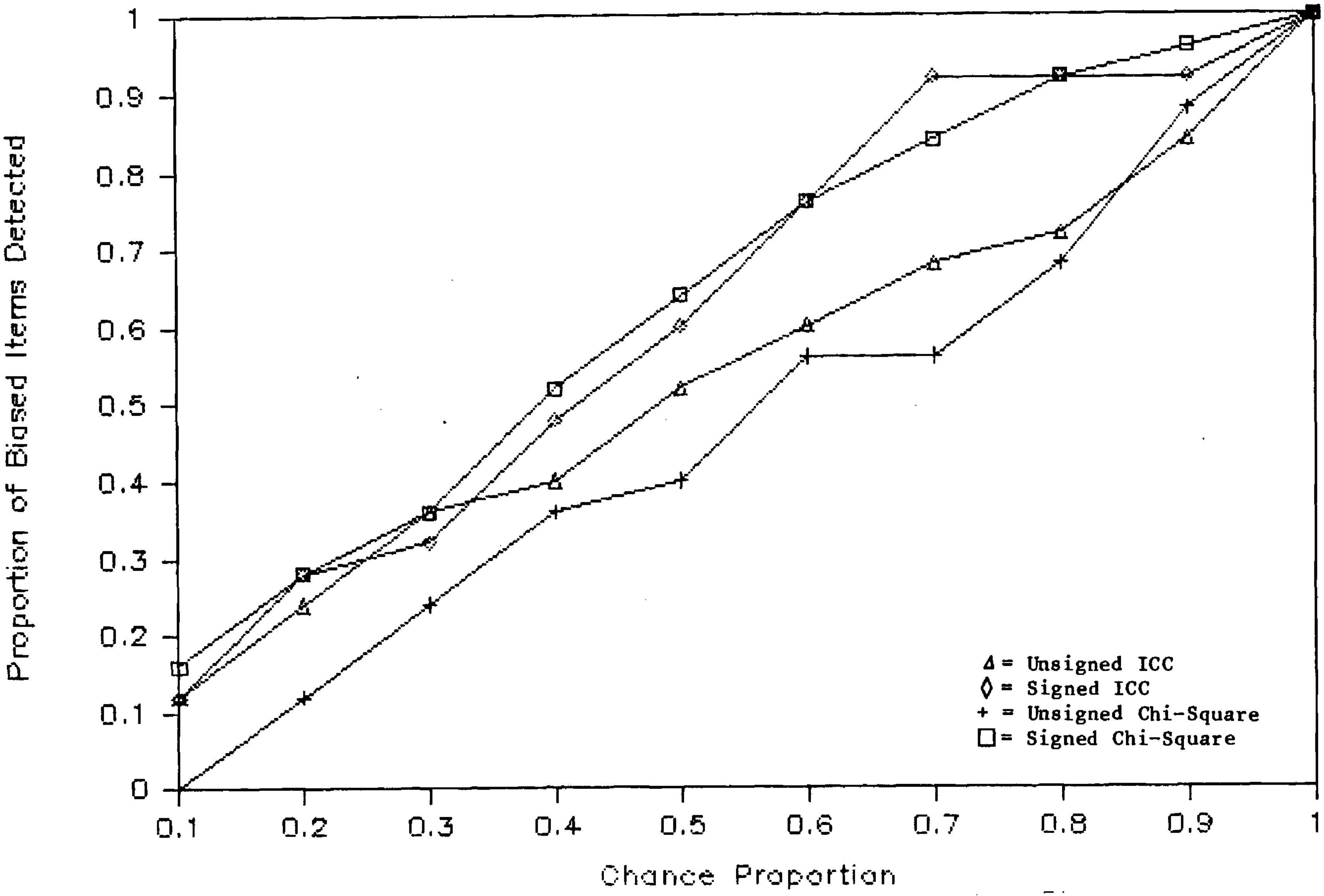

Figure 11. Bias Detection vs. Chance for Case 3

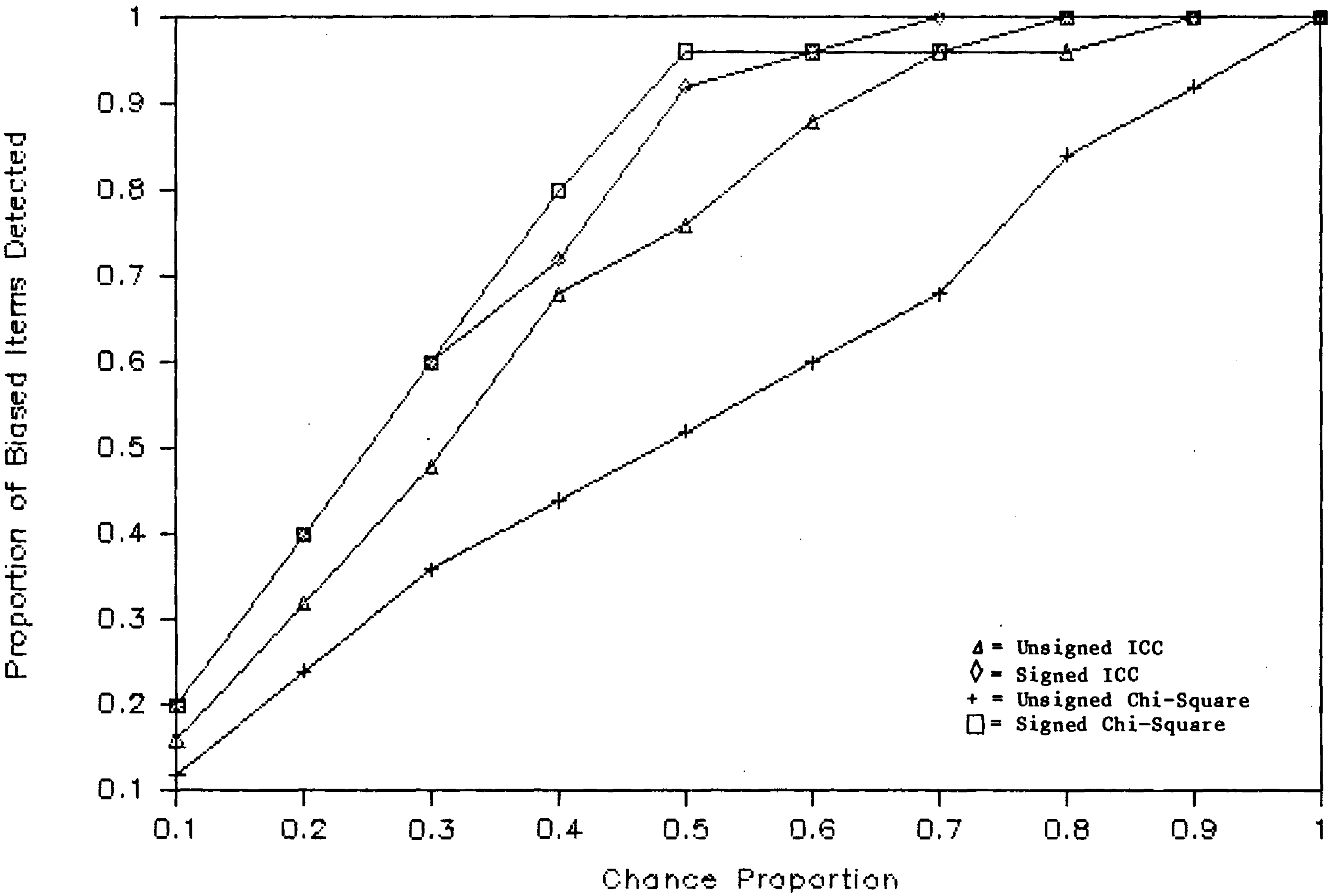

Figure 12. Bias Detection vs. Chance for Case 4