DOCUMENT RESUME

ED 377 223 TM 022 393

AUTHOR Flowers, Claudia P.; Oshima, T. C.
TITLE The Consistency of DIF/DTF across Different Test Administrations: A Multidimensional Perspective.
PUB DATE Apr 94
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ethnic Groups; *Item Bias; Mathematics; *Reliability; *Sex Differences; State Programs; *Test Construction; Testing Programs
IDENTIFIERS *Multidimensionality (Tests)

ABSTRACT

This study was patterned after a previous study by Skaggs and Lissitz (1992) in which inconsistency of differential item functioning (DIF) was reported across test administrations. They suggested multidimensionality of test data as one possible reason for inconsistency. Therefore, in this study, DIF indices which were developed recently with a multidimensional perspective were included. In addition, the consistency of differential test functioning (DTF) was evaluated. DIF/DTF analyses were conducted for both gender and ethnic differences. Ten random samples of 1,000 examinees from each gender and ethnic category were taken from a math basic skills test which was administered in a statewide testing program in two separate years (1984 and 1987). In general, results indicated a more favorable evaluation of the consistency of DIF indices than the Skaggs and Lissitz study. Possible reasons for conflicting conclusions are discussed. (Contains 18 references and 9 tables.) (Author/SLD)

The Consistency of DIF/DTF Across Different Test Administrations: A Multidimensional Perspective

Claudia P. Flowers
Georgia State University
T. C. Oshima
Georgia State University

Running Head: THE CONSISTENCY OF DIF/DTF

Paper Presented at the 1994 Annual Meeting of the American Educational Research Association, New Orleans, LA. Request for information should be sent to T. C. Oshima, Department of Educational Policy Studies, Georgia State University, Atlanta, GA 30303-3083. Bitnet EPSTCO@GSUVM1.GSU.EDU

Abstract

This study was patterned after a previous study by Skaggs and Lissitz (1992) in which inconsistency of differential item functioning (DIF) was reported across test administrations. They suggested multidimensionality of test data as one possible reason for inconsistency. Therefore, in this study, DIF indices which were developed recently with a multidimensional perspective were included. In addition, the consistency of differential test functioning (DTF) was evaluated. DIF/DTF analyses were conducted for both gender and ethnic differences. Ten random samples from each gender and ethnic category were taken from a math basic skills test which was administered in a statewide testing program in two separate years. In general, the results from this study indicated a more favorable evaluation of the consistency of DIF indices than the Skaggs and Lissitz study. Possible reasons for conflicting conclusions are discussed.

The Consistency of DIF/DTF Across Different Test Administrations: A Multidimensional Perspective

Researchers interested in investigating differential item or test functioning (DIF or DTF) continue to look for statistical techniques that are valid and reliable. Many indices have been developed for detecting DIF, but the evaluations of these indices have not always been favorable. Skaggs and Lissitz (1992), for example, reported that the results of DIF analysis were inconsistent and uninterpretable across different test administrations as well as within a test administration. In their study, several DIF methods were applied to a curriculum-based mathematics test for the analysis of DIF among males and females. The consistency of DIF indices across two test administrations (a field-test sample and an operational test sample) was examined. Among the DIF indices they examined, IRT-based sum-of-squares and the Mantel-Haenszel (MH) methods were the most consistent. However, they reported that reliability or agreement of flagged items across different test administrations was modest at best. They recommended a future study with a multidimensional perspective suggesting that the inconsistency might be due to the multidimensionality of test data.

Various reasons have been suggested for the occurrence of false positives (i.e., nonbiased items identified as having DIF) on a test. One of the reasons is model misfit. Researchers have shown that when a test is multidimensional, DIF indices based on

unidimensional models may exhibit DIF due to distributional differences between the two groups of interest (Ackerman, 1992; Oshima & Miller, 1992). Distributional differences, however, may or may not be "biased" depending on the trait(s) a test is intended to measure. Additional traits that the test may measure besides the intended-to-be-measured trait (say, math ability) can be subtle, such as test anxiety, test-wiseness, and speededness.

Recently, new DIF techniques have been proposed which consider the issue of bias in the multidimensional perspective. For example, Raju, van der Linden, and Fleer (1992) proposed an index with which DIF is examined in test data which are meant to be multidimensional. Stout and his colleagues (e.g., Shealy & Stout, 1992) have developed a technique called SIBTEST in which a test developer can choose "valid" items which are unidimensional, then the remaining it .s are tested against those valid items. Additionally, techniques have been developed that examine differential test functioning (DTF) (Raju, et al., 1992;Shealy & Stout, 1992). It is natural to consider "bias" at the test level because bias can be described in the context of multidimensionality which prevails throughout the entire test.

The purpose of this paper was to examine the consistency of those newly developed DIF indices. Consistency was examined across different test administrations (i.e., different random samples from different testing occasions) as well as consistency within the same year administration (i.e., different random samples from the same year). Because DIF analyses from one year

are often used to make decisions about items to include in future tests, it is important to examine the consistency of DIF indices across years. It is also important to examine the consistency within a test administration, because in practice only one sample is taken from each population for a DIF analysis. Additionally, the consistency of DTF was investigated. It is important to note that this study examined only reliability of the indices (i.e., whether or not an item is identified as having DIF consistency across samples) and not validity (i.e., whether or not a truly biased item is identified as having DIF).

Method

Data

The data for this study came from a basic skills mathematics test administered to 10th graders in a state public school system during 1984 and 1987. After excluding examinees who had previously taken the test and those enrolled in special education, 63,406 and 54,605 examinees were in the 1984 and 1987 population data, respectively. Only 2 of the 3 subscales of the test, consisting of 75 items, were used in this study: component operations and problem solving. All 75 items in both the 1984 and 1987 test administrations were identical in content and order of item presentation. All items were multiple choice with 4 options.

DIF analyses focused on differences between gender groups and differences between two ethnic groups. Ten random samples of 1000 examinees were chosen from gender groups (referred to as

Gender 1 and Gender 2) and two ethnic groups (referred to as Ethnic 1 and Ethnic 2) from 1984 and 1987 population data which resulted in a total of 80 samples.

Parameter Estimations and Linking

For the unidimensional solutions, item parameters were estimated using PC-BILOG3 (Mislevy & Bock, 1990). One-parameter (1p), two-parameter (2p), three-parameter (3p), and three-parameter fixed-c (3f) model estimations were performed. No formal goodness of fit analyses of the models were performed and all items were included in this study. Parameter estimates were placed on a common scale using the test characteristic curve method (Stocking & Lord, 1983) utilizing the computer program EQUATE (Baker, 1990).

For the multidimensional solutions, item parameters were estimated using NOHARM (Fraser, 1988). Since the test consisted of two subscales, a two dimensional solution was calculated. Estimated parameters were placed on a common metric using a multidimensional linkage program written in SAS (see Oshima & Davey, 1994 for details).

DIF Indices

For IRT DIF indices, a test item is potentially biased if examinees from one group have a different probability of answering an item correctly than examinees in another group given that the examinees have the same ability level (Hambleton, et al., 1991). The IRT DIF indices examined in this study are listed below:

(a) Signed (SSOS) and Unsigned (USOS) Sums of Squares (Shepard et al., 1984) were included in this study in order to compare the results with Skaggs and Lissitz' (1992) results. USOS is the squared differences of probabilities of a correct response between the two groups given that examinees have the same ability level. The squared differences are summed across all examinees from both groups under study. Because the squared differences will always result in a positive value, USOS will always be positive. The calculation of SSOS is similar to USOS except instead of squaring the differences the absolute value of the difference is multiplied by the original value which results in both positive and negative values. Positive values indicate bias against one group while negative values indicate bias against the other group. Since no distribution was available for SSOS/USOS, criteria for significance were established using the same group baseline comparisons as described in Kim and Cohen (1991). In this method, it was assumed that the areas measured between the two ICCs were normally distributed. One critical value was calculated using a one-tailed .05 level of significance. Areas greater than this critical value were identified as having DIF.

(b) Closed-Interval Signed (CSA) and Unsigned (CUA) (Kim & Cohen, 1991) are calculated by measuring the area between two item characteristic curves (ICC) over the ability range of -4 to 4. CSA and CUA were calculated using a program called IRTDIF (Kim & Cohen, 1991). Because no distribution was available for

closed-interval area measures, the same method used for SSOS/USOS was used to establish cut-off values. CSA and CUA are included in this study because they were not included in Skaggs and Lissitz' study and also considered to provide a contrast with Raju et al.'s index which will be described next.

(c) Raju et al. (1992) proposed a general procedure for assessing DTF and DIF in tests with unidimensional, multidimensional, and polychotomous IRT models. Raju et al. (1992) offered an empirical demonstration of their technique for the unidimensional solution; the multidimensional solution was demonstrated in Oshima, Raju, and Flowers (1993). The program, TBIAS (Raju, et al., 1992), was used to calculate the unidimensional solution (TBU) and the multidimensional solution (TBM). In both solutions, only non-compensatory DIF (NC-DIF) was investigated. All items that exceeded NC-DIF of .006 were identified as displaying DIF.

In addition, the following DIF indices which did not involve IRT item calibrations were examined.

(a) Mantel-Haenszel (MH) Chi-Square (Holland & Thayer, 1986) was calculated in order to compare results of this study against Skaggs and Lissitz' results. MH is the test statistic for the Mantel-Haenszel odds ratio. Items were tested at the .05 level of significance.

(b) Simultaneous Item Bias Test or SIBTEST (Shealy & Stout, 1993) is a procedure that assesses an item or a testlet of items for collective DIF based upon the matching of examinees on the

basis of their score on a specified subtest, called valid subtest. When the matching subtest is asserted to be construct valid, then SIBTEST assesses item and/or test bias. Any value that exceeded the .05 level of significance was identified as displaying DIF.

Results

Table 1 reports the mean, standard deviation, skewness, and kurtosis for the total population, Gender 1, Gender 2, Ethnic 1 and Ethnic 2 for 1984 and 1987 test administrations. Table 2 reports the average mean, standard deviation, skewness, and kurtosis for the 10 random samples drawn from each gender/ethnic group and year (1984 & 1987). There is approximately a .2 standard deviation difference between Gender 1 and Gender 2 and a much larger difference of approximately 1 standard deviation between Ethnic 1 and Ethnic 2. Comparing the samples between years, all groups' scores increased by approximately 2 points, but the differences between the groups remained similar across years.

Insert Tables 1 & 2 about here

The dimensionality of the 80 random samples was examined using DIMTEST (Stout et al., 1992). DIMTEST tests the hypothesis that a set of dichotomously scored items is essentially unidimensional. Of the 20 Gender 1 and 20 Gender 2 samples, 5 out of 20 tests in each gender category rejected the hypotheses

that the tests were unidimensional (.05 level of significance). For Ethnic 1, 17 out of 20 tests violated unidimensionality and Ethnic 2 had only 3 out of 20 samples judged to be multidimensional. Because ethnic comparisons involved more multidimensional data sets, at least for the Ethnic 1 group, multidimensional solutions were performed only between ethnic groups and not gender groups.

Different Test Administrations

Correlation coefficients were calculated as indices of reliability across two test administrations. Each test administration had 10 samples. The average correlation coefficient was obtained for each index by pairing all possible samples from the two test administrations, calculating a correlation coefficient for each pair, and finally calculating the mean of the coefficients over the 100 pairs. The average correlations for all indices are reported in Table 3.

Insert Table 3 about here

The correlation coefficients using SSOS and USOS (the 3p model) were .761 and .655, respectively, for the gender comparisons, and .759 to .613, respectively, for the ethnic comparisons. These values are much higher than those reported in Skaggs and Lissitz (1992). Their coefficients for SSOS and USOS for the 3p model ranged from .36 to .56 with a sample size of 600 or 2000. This suggests that our data exhibited more consistency

than theirs. The CSA/CUA indices showed relatively high consistency. The correlation coefficients ranged from .554 to .876 for the gender comparisons, and .455 to .877 for the ethnic comparisons.

TBU had the highest consistency among the IRT-based indexes examined in this study. The correlation coefficients ranged from .853 to .879 for the gender comparisons, and .637 to .877 for the ethnic comparisons. Interestingly, the correlation coefficients for TBU were fairly constant across different models (i.e., 1p, 2p, 3f, and 3p). Among different models for unidimensional IRT-based indices, the 1p model showed the best consistency. This consistency is possibly due to more stable estimation of item parameters using a lower parameter model. The only index based on multidimensional IRT (TBM) showed the lowest correlation (.498). There appears to be several comparisons that had extremely low correlation coefficients (i.e., less than .10). The outliers are suspected to have occurred due to poor estimation of item parameters by NOHARM.

For indices that did not involve IRT item calibrations, SIBTEST was more consistent than the MH test. For the gender comparisons using SIBTEST, the mean correlation was .863. For the ethnic comparison, it was .809. Even for MH, the correlation coefficients was .833 for the gender comparisons, and .729 for the ethnic comparisons. These values are, again, much higher than those reported in Skaggs and Lissitz (1992) which was reported to range from .30 to .53. In all the indices examined

in the study, the consistency was lower for the ethnic comparisons than for the gender comparisons, suggesting that lower consistency is expected when there is a larger group mean difference.

A more important question concerning the consistency of DIF indices is the agreement of the DIF/NonDIF items (i.e., does the index identify the same items). This was investigated using a two-rater index of agreement, kappa. According to Fleiss, kappas greater than .75 have excellent agreement and kappas lower than .4 have poor agreement (Conger, 1980). All pairwise kappas were calculated between the 1984 and 1987 samples which resulted in 100 kappas per index. The average kappa, standard deviation, minimum and maximum values are reported in Table 4.

Insert Table 4 about here

The trends shown in Table 4 are similar to those shown in Table 3. The average kappa for CSA/CUA ranged from .564 to .222 for gender comparisons and .445 to .291 for ethnic comparisons. TBU performed consistently over all models for the gender comparisons, ranging from .551 to .570. In the ethnic comparisons, the values were slightly lower, ranging from .392 to .502. SIBTEST had agreement values .513 and .499 for gender and ethnic comparisons respectively. Interestingly, MH was the only index that had higher agreement for the ethnic comparisons than the gender comparisons. MH agreement for gender was .484 and

that for ethnic comparisons was .578. All indices had low to moderate agreement across years.

Consistency Within a Single Test Administration

To examine the consistency of each DIF index within a single year test administration, the DIF index was correlated with all other values of that same index within the same year test administration resulting in 45 all possible pairwise correlations for 10 random samples ((10 X 9)/2 = 45). Then the average of these correlations was calculated. Tables 5 and 6 report the average mean, standard deviation, minimum and maximum values of the correlations for gender and ethnic DIF indices.

Insert Tables 5 and 6 about here

The trends in Tables 5 and 6 are similar to those in Table 3 with a very slight increase in correlation in most of the coefficients. These results suggest that the indices are fairly consistent both within a test administration and between two test administrations. Especially, high .80s in correlation coefficients exhibited by TBU and SIBTEST are encouraging.

Agreement of DIF/NonDIF within a single year test administration was investigated using multi-rater kappa suggested by Fleiss (Conger, 1980). Each gender and ethnic index's kappa and average number of items are reported in Tables 7 and 8.

Insert Tables 7 and 8 about here

More items displayed DIF in the ethnic tests verses the gender tests. For ethnic tests the average number of flagged items ranged from 20.0 to 40.9 while gender tests ranged from 10.9 to 36.5. CUA and CSA indices, the methods that required a baseline for establishing cutoff values, tended to identify more items as DIF compared to the other indices except in the 2p models. This is particularly true for the 1p models where the least number of items identified as DIF was 34.5. SIBTEST also identified a large percentage of items as DIF ranging from 28.7 to 35.6. TBU flagged the least items, ranging from 13.6 to 26.0.

No indices' kappas were above .7 which would have shown excellent agreement and several indices exhibited poor agreement. The CUA usually had the poorest agreement (less than .4) except in the 1p model where the kappas are similar to other 1p model indices. All other indices had kappas that ranged from .418 to .681. Among the highest are TBU in all models, SIBTEST, and CSA/CUA in the 1p model. The indices tended to have slightly higher kappa values for gender group tests. This was the same pattern noted in the average correlation results.

Consistency of DTF

Raju et al. (1992) was the only DTF index calculated in this study. 1p, 2p, 3p, and 3f for 1984 and 1987 tests were calculated. For each DTF analysis, a chi-square statistic was calculated to test whether or not an observed DTF is significantly different form zero. Raju et al. recommended that

a single item be identified for removal at a time and the process be continued until the chi-square associated with the revised DTF index becomes nonsignificant. The average chi-square (1000 df), standard deviation, minimum and maximum values are reported in Table 9.

Insert Table 9 about here

The average chi-square value was fairly consistent for each model across years. For example, the 3p model for gender comparisons had a chi-square value of 1060 in 1984 and 1097 in 1987. In both years, only one item at most had to be eliminated to achieve a nonsignificant chi-square value. The item suggested for elimination was not always the same but there were some overlapping items. For example, Items 19 and 63 were flagged at least once for each year.

The consistency within years was examined using the standard deviation and minimum and maximum values. Although several of the models had outliers, for example, in the 1984 2p model for ethnic comparisons there was a standard deviation of 1269.63, which was much higher than other models, the chi-square values were fairly similar from sample to sample. In general, the ethnic DTF showed less consistency than gender DTF.

Conclusions

This study gave a more favorable evaluation of the consistency of DIF indices than Skaggs and Lissitz' (1992) study

did. For the common indices between the two studies (i.e., IRT-based sum of squares methods and the Mantel-Haenszel test), correlation (i.e., reliability) coefficients were much higher in this study than those in Skaggs and Lissitz. Furthermore, other newly developed methods (SIBTEST and TBU) showed even higher consistency than these indices (SSOS/USOS and MH) with correlation coefficients as high as .88. In terms of agreement for flagged items across samples between test administrations and also within a test administration, the Kappa index showed poor to good agreement.

There are several possible reasons for the conflicting conclusions between Skaggs and Lissitz' study and this study. First is related to the equivalence of test forms. In the Skaggs and Lissitz study, two different test forms were used (field-test and operational forms) introducing differences in the order of items. As they pointed out, the position of items within a test can make a difference in estimation of item parameters. In contrast, in this study, both test forms were operational and all the examinees had the exact same tests with items in the same order of presentation. Second is related to the equivalence of samples. As they described, the samples in the two testing occasions might not have been quite equivalent because the field-test sample volunteered to participate in the pilot and were not selected randomly. In our study, 1984 and 1987 samples came from the entire populations of 1984 and 1987, respectively. Thus, the equivalence of samples in terms of their characteristics can be

assumed to be quite similar except a possible year effect.

The two reasons described above indicate important implications in DIF analysis in practice. If in fact the differences in results between the Skaggs and Lissitz study and our study are attributed to the differences in the degree of equivalence in test forms and samples, the importance of constructing and administering a field-test form as similar as possible to an operational form is evident. The difference in test forms (e.g., the order of item presentation) or in samples can certainly introduce additional dimensions which were not present in our study.

Another possible reason for the conflicting conclusions can be simply due to differences in tests and examinees between theirs and ours. For example, more items displayed DIF in this study. Skaggs and Lissitz reported a range of 1 to 14 out of 24 DIF items while DIF items in this study ranged from 14 to 41 out of 75 items. The methodology was another contributor for the difference. In their study, one sample of 1986 was compared against two samples of 1987. In this study, 10 samples of 1984 were compared against 10 samples of 1987, thus allowing less chance of outliers affecting the results. Also note that there was a difference in the sample size. Their sample size ranged from about 600 to 2000 whereas the sample size for this study was always 1000.

The results in this study showed that as the distributional differences between groups being compared increased, the

reliability of the DIF indices tended to decrease. It was particularly interesting that a multidimensional solution (TBM) showed a poor performance in this study. Several reasons are possible. For example, the test was fairly unidimensional contrary to our initial expectation. Therefore, a multidimensional solution was not particularly necessary. When the multidimensional solution was applied regardless, the drawback overweighed the benefit. TBM required multidimensional parameter estimates and a multidimensional linkage method which are both still in the development stage. Recovery of multidimensional item parameters was sometimes problematic with a sample size of 1000. Another interesting observation was that when multidimensionality was present, it often occurred only in one of the two groups of interest. It is not clear which DIF method, if any, can detect dimensionality differences between two groups. Further studies are needed in the area of multidimensionality and DIF.

SIBTEST appeared to be one of the most consistent DIF method and uniformly superior to the Mantel-Haenszel test. Unlike MH, in SIBTEST a suspect item is tested against a valid subtest which is fairly unidimensional. This control of dimensionality may have been the cause of superior performance of SIBTEST over MH. Raju et al.'s TBU was another consistent DIF method. Contrary to CUA/CSA, TBU considers frequency of examinees at a given point of the ability continuum, which may have contributed the enhanced consistency over CUA/CSA.

Finally, results from the consistency of DTF was encouraging. Although not included in this study, DTF using SIBTEST is another index to be studied in the future. Results from this study showed that DIF/DTF indices are not necessarily inconsistent as previous research studies have claimed. As in reliability of the whole test, reliability of DIF indices appear to depend on how these indices are used and should be interpreted accordingly in each context.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Baker, F. B. (1990). EQUATE: Computer program for equating two matrices in item response theory. Madison, WI: University of Wisconsin, Laboratory of Experimental Design.

Conger, A. J. (1980). Integration and generalizations of kappa for multiple raters. Psychological Bulletin, 80, 322-328.

Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Center for Behavioral Studies, The University of New England Armidale, New South Wales, Australia.

Hambleton, R. K., Swaninathan, H., & Rogers, H. J. (1991). Fundamental of item response theory. Newbury Park: SAGE.

Holland, P. W., & Thayer, D. (1986). Differential item performance and the Mantel-Haenszel statistics. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Kim, S., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. Applied Psychological Measurement, 15(3), 269-278.

Mislevy, R. J. & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Oshima, T. C. & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16(3), 237-248.

Oshima, T. C., Raju, N. S., & Flowers, C. P., (1993). Evaluation of DTF and DIF in two-dimensional IRT. Paper presented at the annual meeting of American Educational Research Association, Atlanta.

Oshima, T. C. & Davey, T. (1994). Evaluation of procedures for linking multidimensional item calibrations. Paper to be present at the annual meeting of American Educational Research Association, New Orleans.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992). An internal measure of test bias with application for differential item functioning. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Shealy, R. & Stout, W. (1992). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Office of Navel Research Technical Report.

Shepard, L., Camillie, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Measurement, 9, 93-128.

Skaggs, G. & Lissitz, R. W.(1992). The consistency of detecting item bias across different test administrations: Implications of another failure. Journal of Educational Measurement, 29, 227-242.

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

Stout, William: Namdakumar, Ratna:Junker, Brian:Change, Hua-Hua; Steidinger, Duane (1992). DIMTEST: A Fortran program for assessing dimensionality of binary item responses.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Table 1

Population Raw Score Means, Standard Deviation, Skewness, and Kurtosis for 1984 and 1987

Group	N	Means	SD	Skewness	Kurtosis
		1984 Population			
Total	63,406	52.208	13.420	-.374	-.656
Gender 1	33,017	51.048	13.194	-.282	-.707
Gender 2	30,300	53.475	13.546	-.490	-.546
Ethnic 1	19,665	43.231	12.473	.161	-.575
Ethnic 2	42,488	56.379	11.658	-.599	-.186
		1987 Population			
Total	54,605	53.271	12.881	-.408	-.575
Gender 1	27,786	52.159	12.713	-.310	-.638
Gender 2	26,792	54.426	12.954	-.525	-.450
Ethnic 1	17,059	45.444	12.194	.058	.019
Ethnic 2	36,409	56.903	11.470	-.633	-.122

Table 2

Average Means, Standard Deviation, Skewness, and Kurtosis for the 10 Random Samples Drawn From the Populations

Group	Means	SD	Skewness	Kurtosis
		1984 Samples		
Gender 1	51.066	13.138	-.293	-.662
Gender 2	53.228	13.620	-.474	-.551
Ethnic 1	43.065	12.504	.153	-.525
Ethnic 2	56.483	11.508	-.608	-.119
		1987 Samples		
Gender 1	52.067	12.632	-.287	-.664
Gender 2	54.537	12.852	-.531	-.410
Ethnic 1	45.578	12.137	.064	-.559
Ethnic 2	56.964	11.444	-.636	-.124

Note. All figures are based on sample size of 1000 for each of the 10 replications.

Table 3

A Summary of Correlations for DIF Indices Across Different Test Administrations

	Gender Indices				Ethnic Indices			
Index	Mean	SD	Min	Max	Mean	SD	Min	Max
SSOS-3	.761	.055	.646	.873	.759	.051	.618	.883
USOS-3	.655	.094	.422	.855	.613	.083	.356	.817
TBU-3	.857	.043	.724	.929	.671	.???	.422	.846
CSA-3f	.784	.047	.637	.880	.708	.062	.504	.840
CUA-3f	.608	.085	.347	.748	.455	.091	.193	.649
TBU-3f	.868	.040	.735	.940	.696	.072	.437	.858
CSA-2p	.737	.051	.583	.846	.671	.066	.496	.801
CUA-2p	.554	.087	.274	.715	.468	.073	.263	.610
TBU-2p	.853	.043	.723	.938	.637	.107	.345	.834
CSA-1p	.876	.031	.774	.926	.877	.036	.747	.947
CUA-1p	.742	.065	.526	.855	.754	.054	.582	.871
TBU-1p	.879	.038	.758	.950	.795	.066	.590	.907
SIBTEST	.863	.026	.784	.916	.809	.041	.696	.892
MH	.833	.065	.701	.895	.729	.065	.510	.895
TBM					.498	.177	.055	.864

Table 4

A Summary of Agreement Indices (Kappas) Across Test Administrations

	Gender Indices				Ethnic Indices			
Index	Mean	SD	Min	Max	Mean	SD	Min	Max
SSOS-3p	.337	.110	.081	.640	.439	.104	.232	.714
USOS-3p	.329	.096	.127	.614	.443	.087	.244	.621
TBU-3p	.557	.100	.292	.787	.502	.105	.261	.755
CSA-3f	.373	.073	.202	.574	.312	.086	.117	.515
CUA-3f	.367	.094	.065	.467	.291	.093	.102	.519
TBU-3f	.570	.112	.262	.825	.439	.094	.199	.665
CSA-2p	.364	.085	.191	.536	.326	.097	.101	.624
CUA-2p	.222	.100	.038	.461	.335	.144	.061	.602
TBU-2p	.551	.116	.240	.824	.392	.090	.166	.667
CSA-1p	.560	.090	.361	.785	.445	.100	.211	.627
CUA-1p	.564	.088	.388	.760	.445	.101	.211	.627
TBU-1p	.565	.010	.361	.787	.544	.088	.247	.718
SIBTEST	.513	.075	.377	.707	.499	.088	.284	.755
MH	.484	.092	.282	.658	.578	.089	.373	.755
TBM					.311	.153	-.023	.613

Table 5

A Summary of Correlations for DIF Indices Within the Same Year Test Administration for Gender Groups

	1984 Samples				1987 Samples			
Index	Mean	SD	Min	Max	Mean	SD	Min	Max
SSOS-3p	.766	.057	.648	.875	.777	.046	.679	.847
USOS-3p	.668	.088	.451	.833	.657	.073	.500	.773
TBU-3p	.877	.045	.760	.950	.849	.034	.764	.908
CSA-3f	.791	.046	.697	.892	.807	.040	.713	.890
CUA-3f	.637	.082	.447	.819	.609	.076	.414	.770
TBU-3f	.886	.047	.776	.960	.863	.027	.784	.911
CSA-2p	.741	.052	.649	.846	.751	.043	.644	.866
CUA-2p	.593	.079	.415	.758	.554	.088	.321	.746
TBU-2p	.876	.046	.788	.938	.593	.162	.303	.908
CSA-1p	.885	.031	.826	.943	.889	.028	.811	.935
CUA-1p	.766	.063	.636	.901	.760	.052	.620	.849
TBU-1p	.891	.044	.811	.958	.875	.032	.791	.931
SIBTEST	.872	.026	.799	.921	.875	.019	.835	.913
MH	.850	.052	.736	.925	.837	.042	.706	.906

Table 6

A Summary of Correlations for DIF Indices Within the Same Year Test Administration for Ethnic Groups

	1984 Samples				1987 Samples			
Index	Mean	SD	Min	Max	Mean	SD	Min	Max
SSOS-3p	.777	.046	.676	.859	.793	.048	.675	.878
USOS-3p	.637	.070	.472	.787	.669	.081	.457	.834
TBU-3p	.674	.067	.527	.827	.740	.055	.599	.845
CSA-3f	.769	.057	.619	.886	.737	.060	.560	.827
CUA-3f	.530	.093	.313	.730	.517	.089	.204	.719
TBU-3f	.705	.064	.542	.846	.757	.050	.625	.838
CSA-2p	.744	.062	.608	.861	.707	.070	.514	.828
CUA-2p	.553	.078	.362	.682	.501	.101	.252	.662
TBU-2p	.690	.066	.535	.801	.749	.065	.602	.885
CSA-1p	.902	.026	.826	.940	.903	.013	.876	.929
CUA-1p	.801	.046	.666	.859	.780	.029	.706	.856
TBU-1p	.832	.066	.709	.883	.841	.041	.731	.922
SIBTEST	.822	.030	.762	.895	.826	.037	.744	.826
MH	.729	.069	.593	.864	.750	.063	.604	.888
TBM	.441	.207	.038	.871	.624	.165	.209	.930

Table 7

Kappa and Average Number of Items Identified as DIF for Gender Groups

	1984 Samples		1987 Samples	
Index	Kappa	Average Number of Biased Items	Kappa	Average Number of Biased Items
SSOS-3p	.451	19.2	.418	25.1
USOS-3p	.353	17.4	.357	24.1
TBU-3p	.586	15.1	.570	15.1
CSA-3f	.536	22.2	.487	31.3
CUA-3f	.379	18.0	.332	27.2
TBU-3f	.597	13.6	.570	13.7
CSA-2p	.448	16.8	.435	24.7
CUA-2p	.309	10.9	.300	25.4
TBU-2p	.605	13.6	.558	14.5
CSA-1p	.653	36.5	.681	34.5
CUA-1p	.571	36.5	.617	34.5
TBU-1p	.567	14.7	.571	15.4
SIBTEST	.599	35.6	.609	33.8
MH	.536	27.0	.561	33.8

Table 8

Kappa and Average Number of Items Identified as DIF for Ethnic Groups

	1984 Samples		1987 Samples	
Index	Kappa	Average Number of Biased Items	Kappa	Average Number of Biased Items
SSOS-3p	.497	25.7	.543	24.5
USOS-3p	.474	24.6	.472	22.3
TBU-3p	.512	24.3	.511	24.3
CSA-3f	.476	34.0	.479	21.7
CUA-3f	.358	35.3	.391	23.3
TBU-3f	.499	21.6	.494	20.2
CSA-2p	.492	30.0	.455	22.7
CUA-2p	.396	34.2	.357	26.6
TBU-2p	.513	26.0	.492	20.0
CSA-1p	.607	36.4	.573	40.9
CUA-1p	.518	37.1	.450	40.4
TBU-1p	.557	21.2	.579	21.5
SIBTEST	.563	28.7	.542	29.1
MH	.550	25.8	.578	25.7
TBM	.176	30.3	.361	32.8

Table 9

A Summary of Chi-Square Values for Raju et al.'s DTF

	Gender Comparisons							
	1984				1987			
Model	Chi-Sq	SD	Min	Max	Chi-Sq	SD	Min	Max
3p	1060.00	46.46	1004	1143	1097.00	58.11	1002	1187
3f	1085.50	53.63	1002	1162	1098.40	42.62	1000	1135
2p	1125.20	162.72	1000	1537	1091.50	79.35	1001	1243
1p	2533.30	121.80	2350	2629	2984.70	144.36	2815	3233
	Ethnic Comparisons							
3p	1508.20	147.86	1332	1697	1557.00	327.19	1205	2247
3f	1351.40	430.58	1002	2521	1405.20	149.58	1070	1556
2p	2166.00	1269.63	1249	5708	1286.70	173.45	1045	1552
1p	1243.10	34.66	1195	1299	1569.40	43.46	1528	1667