DOCUMENT RESUME

ED 300 399 TM 012 374

AUTHOR Rogers, H. Jane; Hambleton, Ronald K.

TITLE Evaluation of Computer Simulated Baseline Statistics for Use in Item Bias Studies. Revised .

PUB DATE Apr 87

NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). For original document of which this is a revision, see ED 287 870.

PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Computer Simulation; Evaluation Methods; *Item Analysis; Latent Trait Theory; Secondary Education; *Statistical Analysis; Statistical Bias; *Test Bias; Test Construction; Test Interpretation; Test Items

IDENTIFIERS *Baseline Statistics; *Item Bias Statistics

ABSTRACT

Although item bias statistics are widely recommended for use in test development and test analysis work, problems arise in their interpretation. The purpose of the present research was to evaluate the validity of logistic test models and computer simulation methods for providing a frame of reference for item bias statistic interpretations. Specifically, the intent was to produce simulated sampling distributions of item bias statistics under the hypothesis of no bias for use in determining cut-off points to provide guidelines for interpreting item bias statistics obtained with actual test data. The test data used were the item scores of 207 white and 730 black Cleveland (Illinois) ninth graders to the 75 items on the 1985 Cleveland Reading Competency Test. The area, root mean squared difference, and Mantel-Haenszel methods were used to statistically analyze the data. The results support the basic data simulation approach used in this study. Real and simulated distribution for three item bias statistics when bias was not present were very similar and the minor differences that were found between the distributions had little effect on the interpretations of item bias statistics obtained with actual test data. Seven steps for applying the method of computer-simulated baseline statistics in test development settings are outlined. One data table and four graphs conclude the document. (Author/SLD)

# Evaluation of Computer Simulated Baseline Statistics for Use in Item Bias Studies

H. Jane Rogers and Ronald K. Hambleton
University of Massachusetts at Amherst

5/30/88

# Evaluation of Computer Simulated Baseline Statistics for Use in Item Bias Studies[1,2]

H. Jane Rogers and Ronald K. Hambleton
University of Massachusetts at Amherst

## Abstract

Though item bias statistics are widely recommended for use in test development and test analysis work, problems arise in their interpretation. The purpose of the present research was to evaluate the validity of logistic test models and computer simulation methods for providing a frame of reference for item bias statistic interpretations. Specifically, the intent was to produce simulated sampling distributions of item bias statistics under the hypothesis of no bias for use in determining cut-off points to provide guidelines for interpreting item bias statistics obtained with actual test data.

The results provided support for the basic data simulation approach used in the study. Real and simulated distributions for three item bias statistics when bias was not present were very similar and the minor differences that were found between the distributions had little effect on the interpretations of item bias statistics obtained with actual test data. Seven steps for applying the method of computer-simulated baseline statistics in test development settings were outlined in the paper.

The great public concern in this country over unfairness or bias in testing has resulted in substantial numbers of research studies that have described and evaluated new methods for identifying potentially biased test items (Berk, 1982; Shepard, Camilli and Averill, 1981; Shepard, Camilli and Williams 1985). Most of the new methods based upon use of item response models and related procedures involve the calculation of statistics which are unfamiliar to test developers (e.g., weighted b value differences, area between two item characteristic curves, sum of squared differences between two item characteristic curves).

One problem that has arisen in test development work concerns the interpretations of these new item bias statistics. Certainly the statistics, whatever their interpretation, can be used to rank-order test items to identify the items of most and least concern. As test developers often want to sort test items into ordered categories (e.g., "must be very carefully reviewed", "may need revision", "should be acceptable"), critical values or cut-off points for classifying the item bias statistics would be useful. The advantage of a classificatory approach, as opposed to an approach based upon item rankings, is that the number of potentially biased items does not need to be specified in advance of the analysis. Thus the number of items identified as potentially biased would depend on the dataset. Of course the main difficulty in placing items into categories is determining a frame of reference and subsequently cut-off scores for interpreting the IRT item bias statistics of interest.

The main purpose of the present research was to evaluate the validity of logistic test models and computer simulation methods for generating sampling distributions of item bias statistics under the hypothesis of no item bias. These distributions are intended for use in setting cut-off points to provide baselines for interpreting item bias statistics. A secondary purpose was to highlight the use of the methods in an item bias study.

This study was prompted by some earlier research by Hambleton, Rogers, and Arrasmith (1986). These authors carried out a similar study obtaining baselines from the analysis of real data provided by two randomly equivalent majority samples and by two randomly equivalent minority samples. Although meaningful baseline results are available by conducting item bias studies on randomly equivalent samples, the disadvantage of this approach is that the important comparisons between the majority and minority groups are carried out with sample sizes half that of those sample sizes that were actually available.

Reduction of sample sizes by 50% to obtain baseline information is a high price to pay when initial sample sizes are often not very large. Small sample item bias studies are especially problematic when IRT methods are used (Hoover and Kolen, 1984). Hambleton et al. (1986) also showed that logistic models could be used to provide simulated results to serve as a baseline for interpreting item bias statistics. It was clear, however, that more research was needed to strengthen their conclusion.

Another way that item bias baseline statistics might be compiled is by combining the majority and minority groups of interest and then by conducting an item bias investigation using two randomly equivalent samples drawn from the combined sample (Shepard, Camilli, and Williams, 1984; Wilson-Burt, Fitzmartin, and Skaggs, 1986). As item bias should not be present in two randomly equivalent groups, the distribution of item bias statistics obtained in two randomly equivalent groups could serve as a basis for setting cut-off scores for interpreting item bias statistics in the majority and minority samples.

The main shortcoming of this approach -- a shortcoming of the early Hambleton et al. (1986) work too -- is that any difference in the ability distributions between the majority and minority groups is not reflected in the two randomly equivalent samples used to obtain the baseline statistics. As group ability distributions can influence the quality of item bias statistics (e.g., Shepard et al., 1984; Wilson-Burt et al., 1986), failure to incorporate this information in the analysis could reduce the usefulness of the distribution of item bias statistics derived from the two randomly equivalent samples. One solution that is sometimes applied when the majority group is large involves selecting an examinee sample from the majority group to approximate the distribution of scores in the minority group (e.g., Shepard et al., 1984). On the other hand, such ability differences and other unique features of the majority and minority samples can be incorporated into a computer-simulated item bias analysis regardless of the available sample sizes. For this reason, the current research centered on the potential value of computer-simulation techniques for providing the desired baseline distributions.

## Method

### Choice of Item Bias Statistics

Three popular item bias statistics were chosen for the investigation: area method, root mean squared difference method, and the Mantel-Haenszel method. The choice of statistics was not of paramount importance to the study, as the purpose was to investigate the usefulness of simulated baseline distributions for the statistics rather than the value of the statistics themselves. The methodology to be proposed could be used with any of the other popular bias statistics, such as the pseudo-IRT, the full chi-square, or the residualized delta, for which the distributional properties are either known only approximately or not at all. Although the Mantel-Haenszel statistic does have a known distribution theoretically, it was included in the study because of the current interest in it, and because of quickness and ease of calculation.

Area Method. In the Area Method, or Total Area Method as it is sometimes called, the area between item characteristic curves for the same item obtained in the majority and minority groups over a specified interval on the ability scale (-3 to +3, in this study) is used as an estimate of item bias (Rudner, Getson, and Knight, 1980). An item is labeled as "potentially biased" when the area between the two curves is large.

Root Mean Squared Difference Method. In applying this method (Linn, Levine, Hastings, and Wardrop, 1981), one calculates the squared difference between the majority and minority item characteristic curves at fixed intervals (usually .01). These squared differences are calculated over the interval on the ability scale which is of interest. Finally, an average of the squared differences is calculated and the square root of the average is taken. Again, large-valued statistics reflect substantial differences between item characteristic curves. Consequently, items associated with large-valued statistics are labeled as "potentially biased."

Mantel-Haenszel Method. The Mantel-Haenszel method has generated considerable interest among test developers in recent years because it appears to provide a quick, cheap, and valid indicator of item bias (Holland and Thayer, 1986). Unlike the other two methods, this method does not involve the application of item response theory (IRT) models and principles. In essence, the method first matches examinees on a criterion variable, often the overall test score because of convenience. The ratio of the odds for success of the majority and minority group members are calculated in each score group of interest (with $\underline{n}$ items, with $\underline{n}+1$ possible score groups). Each ratio is weighted by the sample size in the score group and then the ratios for the (up to) $\underline{n}+1$ score groups are combined to obtain the Mantel-Haenszel statistic. When the odds for success on an item in the majority and minority groups among examinees of the same ability level are substantially different, item bias is suspected. The advantage of this method over the other two previously described ones is that there is an associated statistical test with a known sampling distribution (chi-square with one degree of freedom). Thus meaningful cutoff scores can be established. This statistic was considered because of the substantial interest in its use in item bias work.

Description of the Test Data and Examinee Sample

The test data used in the study were the item scores of 937 Cleveland ninth-grade students to 75 items on the 1985 Cleveland Reading Competency Test (Cleveland Public Schools, 1985). In the total sample, 207 Whites and 730 Blacks were present, of whom 451 were males and 486 were females. Because of the very small number of whites in the sample, only a sex bias study was completed.

Generation of Simulated Examinee Item Scores

Basically, the approach was to simulate examinee item score data that reflected as closely as possible the actual examinee and item data of interest without any item bias. Item parameter and ability parameter estimates obtained from the combined group three-parameter logistic model analysis were treated as "true values" and then a simulated set of item scores for the 937 examinees was generated by using the three-parameter logistic model (Hambleton and Rovinelli, 1973).

With known ability, θ, and model parameters for item i, denoted ai, bi, ci, the probability of the examinee answering the item correctly was assumed to be given by the three-parameter logistic model:

$$P_i(\theta) = c_i + (1-c_i)\ [1 + e^{-Da_i(\theta-b_i)}]^{-1}$$

With $P_i(\theta)$ in hand, an item score, 0 or 1, was obtained by first choosing a random number from a uniform distribution on the interval [0, 1]. If the random number chosen was less than or equal to $P_i(\theta)$, which happens $P_i(\theta)$ of the time, the examinee was scored 1; otherwise the examinee was scored 0. This process was repeated for each of the 75 items for the first examinee using the item parameter estimates obtained from the analysis of the 937 examinees on the 75 item test. Then, the ability score for the second examinee was substituted for the first examinee in Equation [1], and the process of generating a vector of item scores was repeated. This process was continued until 937 vectors of item scores were generated.

The final product was a complete set of item scores for the 937 examinees on the 75 items that were manifested from the three-parameter logistic model. The simulated item scores were generated to be consistent with the item and ability parameter estimates obtained with the real data, but without bias. There was no bias because male and female item scores were generated from a common set of three-parameter item characteristic curves. Any differences in ability scores between the majority and minority groups were retained because the ability estimates obtained from the analysis of the real data were used in the simulations.

A parallel set of item bias analyses was carried out on the real and simulated data. Differences in the distributions of item bias statistics would arise if bias were present in the real data, as in all other respects, the datasets were equivalent, if one assumes, of course, that the three-parameter logistic model provided an appropriate fit to the real data. For this reason, the fit of the three-parameter logistic model to the test data was checked carefully (Hambleton and Rogers, in press; Hambleton and Swaminathan, 1985).

## Procedure

With the actual and simulated test data in hand, three sets of analyses were completed. The first analysis was intended to evaluate the merits of computer simulated baseline sampling distributions of item bias statistics. This analysis involved the comparison of distributions of item bias statistics obtained in randomly equivalent groups (no bias present) through using the real data and the simulated data. In this study, the available samples (real and simulated) were halved in the analyses to provide a basis for evaluating the merits of the chosen simulation methods.

The second analysis was intended to address the comparative effects of employing simulated rather than real sampling distributions in setting cut-off scores. This analysis involved (a) setting cut-off scores with both the real and simulated sampling distributions of item bias statistics obtained under the true hypothesis of no bias and (b) comparing the effect of the different cut-off scores on the number of items labelled "potentially biased" in a sex bias study.

The third and final analysis was an application of the new method in a male-female item bias study. In this analysis, the purpose was to highlight how the method can work in practice.

The specific steps in the procedure were as follows:

1. The real dataset was split into 4 subgroups, two male and two female, denoted $M_1$, $M_2$, $F_1$, and $F_2$. The $M_1$ and $M_2$, and the $F_1$ and $F_2$ subgroups were randomly equivalent. Subgroups were formed so that an item bias study in the two randomly equivalent male samples and in the two randomly equivalent female samples could be achieved. The distribution of these item bias statistics (no bias present) provided a basis for evaluating the distribution generated from the simulated test data. Next, the simulated test data were also divided into four subgroups: $M_1$, $M_2$, $F_1$, and $F_2$. In this way, item bias statistics in the $M_1$ and $F_1$ and in the $M_2$ and $F_2$ samples in

the simulated data could be calculated for the purpose of producing a sampling distribution of each item bias statistic of interest under the hypothesis of no bias. Both $M_1$ and $F_1$ and $M_2$ and $F_2$ comparisons were preferred to the corresponding $M_1$ and $M_2$ and the $F_1$ and $F_2$ comparisons because the former subgroups reflected any real ability differences in the male and female samples, whereas the latter subgroups did not.

2. Separate modified three-parameter model analyses of the $M_1$, $M_2$, $F_1$, and $F_2$ real and simulated data were carried out. The $c$ parameter was fixed at a value of .20. Eight IRT analyses, in all, were completed. Ability estimates obtained from the combined group analysis were also fixed in these analyses.

3. After the necessary data rescalings, two of the item bias statistics of interest -- Area and Root Mean Squared Difference -- were calculated for the group comparisons listed below:

Real Data

a. $M_1$ vs $F_1$

b. $M_2$ vs $F_2$ (this analysis served as a replication of the study with the $M_1$ and $F_1$ samples)

c. $M_1$ vs $M_2$

d. $F_1$ vs $F_2$

e. $M$ vs $F$

Simulated Data

f. $M_1$ vs $F_1$

g. $M_2$ vs $F_2$

h. $M$ vs $F$

The Mantel-Haenszel statistics were calculated using the item response data provided at step 1.

4. For each item bias statistic, the following distributions were obtained:

Real Data

a. The combined distribution of $\underline{M}_1$ vs $\underline{M}_2$ and $\underline{F}_1$ vs $\underline{F}_2$ item bias statistics. (This distribution served as the baseline for interpreting the real item bias statistics obtained from the $\underline{M}_1$ vs $\underline{F}_1$ and $\underline{M}_2$ vs $\underline{F}_2$ comparisons.)

b. The distributions of the $\underline{M}_1$ vs $\underline{F}_1$ and of the $\underline{M}_2$ vs $\underline{F}_2$ item bias statistics. (The $M_2$ vs $F_2$ comparison served as a replication of the $M_1$ vs $F_1$ comparison.)

Simulated Data

c. The combined distribution of $\underline{M}_1$ vs $\underline{F}_1$ and of $\underline{M}_2$ vs $\underline{F}_2$ item bias statistics. (This distribution served as the alternate baseline for interpreting the real item bias statistics obtained from the $\underline{M}_1$ vs $\underline{F}_1$, and $\underline{M}_2$ vs $\underline{F}_2$ groups.) This distribution was compared to 4(a) cited previously to assess the viability of the computer-generated sampling distributions of item bias statistics.

5. The distributions obtained in step 4 (except for the real $\underline{M}$ vs $\underline{F}$ comparison) were smoothed by the method of "weighted rolling averages" (Kendall and Stuart, 1968) to remove some of the minor irregularities in the distributions.

6. The cut-off score corresponding to the .05 level cf significance for each distribution (real and simulated) generated under the hypothesis of no bias was determined.

7. The cut-off scores obtained at step 6 were applied to the real item bias statistics to compare their effects.

In a final phase of the research, the IRT computer simulation method was used to provide a baseline distribution for interpreting item bias statistics obtained in the full male and female samples.

## Results

### Model-Data Fit

The results from this study would have been meaningless unless the three-parameter logistic model had at least provided an adequate accounting of the actual item score data. Fortunately, the model fit the test data well. The average residual (actual performance-expected performance assuming model-data fit) was .01. This average was based on 12 comparisons (at ability levels -2.75, -2.25, ..., 2.75) of the observed and expected performance for each of the 75 items in the test. Clearly, there was no overall bias in the fit of the item and ability parameter estimates to the test data. The average residual calculated at each of the same ability levels across the 75 items was also very small. It exceeded a value of .05 at four ability levels, -2.75, -2.25, -1.75, and 2.75 where the combined examinee sample was only 71 (about 7.5% of the total sample). In sum, the goodness-of-fit results indicated a close fit between the three-parameter logistic model and the actual test data.

Comparison of the Real and Simulated Null Distributions

Figures 1, 2, and 3 provide the smoothed distributions under the hypothesis of no bias for the three item bias statistics with both real and simulated data. The results were clear: There was very little difference between the sampling distributions of the item bias statistics generated with real and simulated data. The maximum difference in the sampling distributions with real and simulated data was 7.8%. Also, the largest differences were always observed in the lower halves of the sampling distributions where the consequences of differences between the distributions on the determination of cut-off values were small.

---

Insert Figures 1, 2, and 3 about here

---

Effect of Choice of Sampling Distribution

Perhaps the best way to judge the effects of choosing the simulated over the real distributions of item bias statistics under the hypothesis of no bias is in terms of the practical consequences of using the cut-off scores obtained from the two distributions. Table 1 provides the .05 cut-off score for the real and simulated distributions for each item bias statistic under the hypothesis of no bias. These cut-off scores corresponded to the 95th percentile of the distribution of statistics in each case. These cut-off scores were then applied to the $M_1$ vs. $F_1$ and to the $M_2$ vs. $F_2$ real item bias data.

---

Insert Table 1 about here

---

Table 1 shows that there were differences in the values of cut-off scores obtained with the real and simulated distributions. These differences influenced the numbers of test items identified at the .05 level, though the influence of choice of distribution appeared to be small. Across six comparisons, the average difference was three items. In view of the close similarity in the distributions as reflected in Figures 1, 2, and 3, it is likely that the differences reflected, to a great extent, the instability in determining the 95th percentile because of the very limited amounts of data in the tails of the distributions. Smoothing the simulated distributions was helpful, but basically the problem remained: there was a limited number of data points in the tails of the distributions. In addition, some differences in the results were expected because the simulated distributions reflected the ability distribution differences in the male and female samples to a greater degree than the real null distributions under the hypothesis of no bias.

An Example

Though samples of (approximately) 450 males and females were available for the research investigation, it was necessary to divide each sample in half so that various comparisons of results could be made to evaluate the merits of the computer simulation. In practice, a test developer would carry out the item bias study with the full set of available data. Figure 4 highlights the results of an item bias investigation (using the Total Area Statistic) with the full male and female samples, and the smoothed computer-simulated distribution of the total arca item bias statistics without any bias. The .05 level of significance was chosen to identify items in need of careful review. The number of items identified was eight. Similar analyses were carried out with the other two bias statistics of interest in this study. Six items were identified with the Root Mean Squared Difference Method; while seven items were identified with the Mantel-Haenszel method.

--------------------------------------------------------------

Insert Figure 4 about here

--------------------------------------------------------------

## Conclusions

The main results of this study reported in Figures 1 to 3 provided support for the use of simulated data to establish critical values for IRT item bias statistics. When the test data fit the model chosen, use of the IRT parameter estimates to generate data allows the test developer to simulate samples closely resembling the original data but under conditions of no bias. Though the results in Figures 1 to 3 do not provide evidence of the importance of retaining ability differences in the simulations of majority and minority group performance, nevertheless, preserving these differences to enhance the validity of the simulated sampling distributions seems desirable. Given the practical limitations of IRT parameter estimation, particularly in relatively small samples, retaining these ability distribution differences may be important, as they may affect the IRT item bias statistics. When randomly equivalent samples of the real data are used to establish cutoff values for the bias statistics, this consideration is not taken into account. Hence simulating the ability differences under conditions of no bias probably allows the investigators to set more realistic cutoff values for the bias statistics.

In the present study, taking ability distribution differences, though slight, into account, produced higher cutoff values with IRT-based methods than were obtained from using random samples of the real data. The result was the flagging of fewer items as biased. Given that the groups were males and females, and that no substantial bias was expected, the direction of the observed differences supports the use of simulated data to establish cut-off points for the IRT item bias statistics.

The lack of agreement observed between the two replications of the bias analysis in the real data (as revealed in Table 1) highlights the problem of using IRT methods in small samples. Substantially better results should be obtainable with larger sample sizes. But with small samples, researchers should be cautioned against using any firm cut-off score for the bias statistics. In the small sample case it is recommended that the simulated data baseline be used more to give a sense of what is extreme in the values of the bias statistics than to label an item as potentially biased or not. Smoothing distributions definitely reduced the problem of unstable cut-off points; using larger samples would be very helpful too.

The results for the Mantel-Haenszel statistic suggest that although data can be generated which will return IRT parameter estimates similar to those obtained from the real data, it is more difficult to generate response patterns that closely resemble the real data. Hence, the method proposed in this paper of simulating data to establish baseline values may not be useful for bias statistics that are not derived from IRT models.

In summary, application of the IRT computer simulation method for generating baseline distributions of item bias statistics is as follows:

1. Choose an IRT model and estimate item and ability parameters for the total group of examinees. Assess model-data fit. Continue with the method if the model-data fit is acceptable. Otherwise choose a more general IRT model to fit the data better. Items which are suspected of being biased can be removed from the analysis at this step. Removal of items does not seem necessary unless the number of items suspected of being biased is a significant portion of the total number of items in the test (e.g., 10% or more).
2. Treat the item and ability parameter estimates as "true" values and generate a new set of examinee item scores by using the logistic model of choice in step 1 (e.g., Hambleton and Rovinelli, 1973).
3. Split the simulated examinee item scores into the majority and minority groups of interest and re-estimate the item parameters, while treating ability scores obtained at step 1

as fixed. (Fixing the ability scores serves two purposes: (a) item parameter estimation time is reduced substantially and (b) scaling problems with the data are considerably reduced.)

4. Choose the IRT item bias statistic (or statistics) of interest and carry out the necessary calculations on the ICCs and ability estimates for the simulated majority and minority test data.

5. Produce the sampling distribution of the item bias statistics obtained from the simulated data and smooth the distribution of resulting item bias statistics to remove some of the instability in determining cut-off scores. Determine the cut-off value corresponding to the 95th percentile (and/or other cut-off values of interest).

6. Repeat steps 3 and 4 with the real test data.

7. Interpret the item bias statistics obtained with the real test data at step 4 by using the cut-off values obtained from the simulated test data at step [illegible]

Test developers who carry out these seven steps will be able to interpret their item bias statistics more meaningfully due to the availability of information about the distribution of the item bias statistics when no bias is present.

## References

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Cleveland Public Schools. (1965). Cleveland Reading Competency Test. Cleveland, OH: Author.

Hambleton, R.K., and Rogers, H.J. (in press). Promising directions for assessing item response model fit to test data. Applied Psychological Measurement.

Hambleton, R.K., Rogers, H.J., and Arrasmith, D. (1986, August) Identifying potentially biased test items: A comparison of the Mantel-Haenszel statistic and several item response theory methods. Paper presented at the annual meeting of APA, Washington.

Hambleton, R.K., and Rovinelli, R.J. (1973). A Fortan IV program for generating examinee response data for logistic test models. Behavioral Science, 18, 74.

Hambleton, R.K., and Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

Holland, P.W., and Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Technical Report No. 86-31. Princeton, NJ: Educational Testing Service.

Hoover, H.D., and Kolen, M.J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Kendall, M.G., and Stuart, A. (1968). The advanced theory of statistics, Volume 3. New York: Hafner Publishing Co.

Linn, R.L., Levine, M.V., Hastings, C.N., and Wardrop, J.L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

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

Shepard, L., Camilli, G., and 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.

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

Shepard, L., Camilli, G., and Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Wilson-Burt, C., Fitzmartin, R.D., and Skaggs, G. (1986, April). Baseline strategies in evaluating IRT item bias indices. Paper presented at the annual meeting of AERA, San Francisco.

## Footnotes

[1] Laboratory of Psychometric and Evaluative Research Report No. 162. Amherst, MA: School of Education, University of Massachusetts.

[2] A paper presented at the annual meeting of AERA, Washington, 1987.

Table 1

Effects of the Choice of Distribution (Real or Simulated) on the Determination of Cut-off Scores and Identification of Potentially Biased Test Items

| | Real Null Distribution | | Simulated Null Distribution | | Difference |
|---|---|---|---|---|---|
| Bias Statistic | Critical Value[1] | Biased Items[2] | Critical Value | Biased Items | |
| Area | .544 | 4<br>(11) | .659 | 1<br>(6) | 3<br>(5) |
| Root Mean Squared Difference | .113 | 4<br>(10) | .134 | 3<br>(3) | 1<br>(7) |
| Mantel-Haenszel | 3.42 | 6<br>(19) | 3.03 | 6<br>(21) | 0<br>(2) |

[1] At the .05 level.

[2] The numbers in brackets correspond to the numbers of test items identified as potentially biased in a replication of the study with the second male and female samples.

## Figure Captions

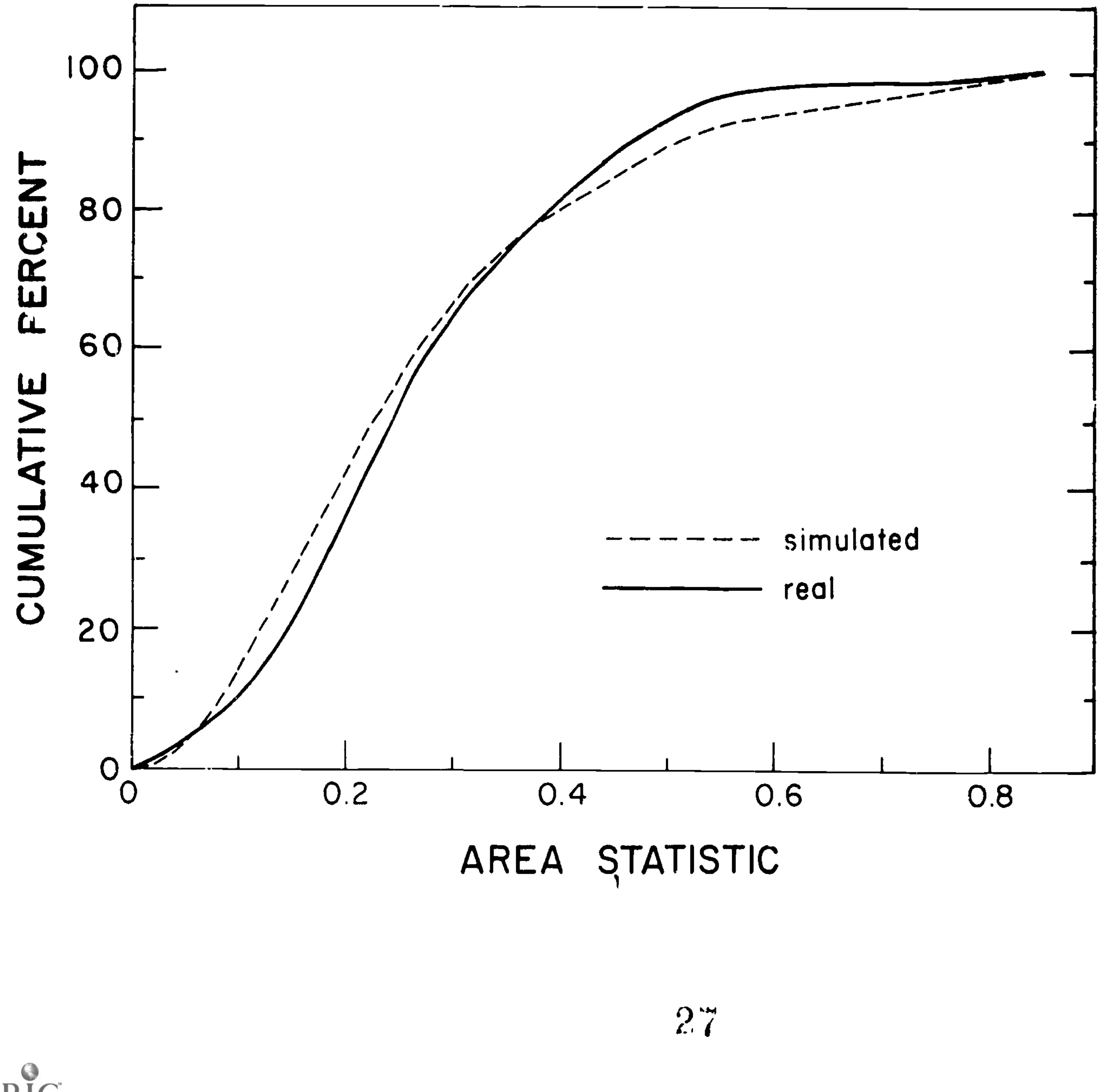
Figure 1. A comparison of the simulated and real sampling distributions of the Item Area Statistics under the hypothesis of no bias.

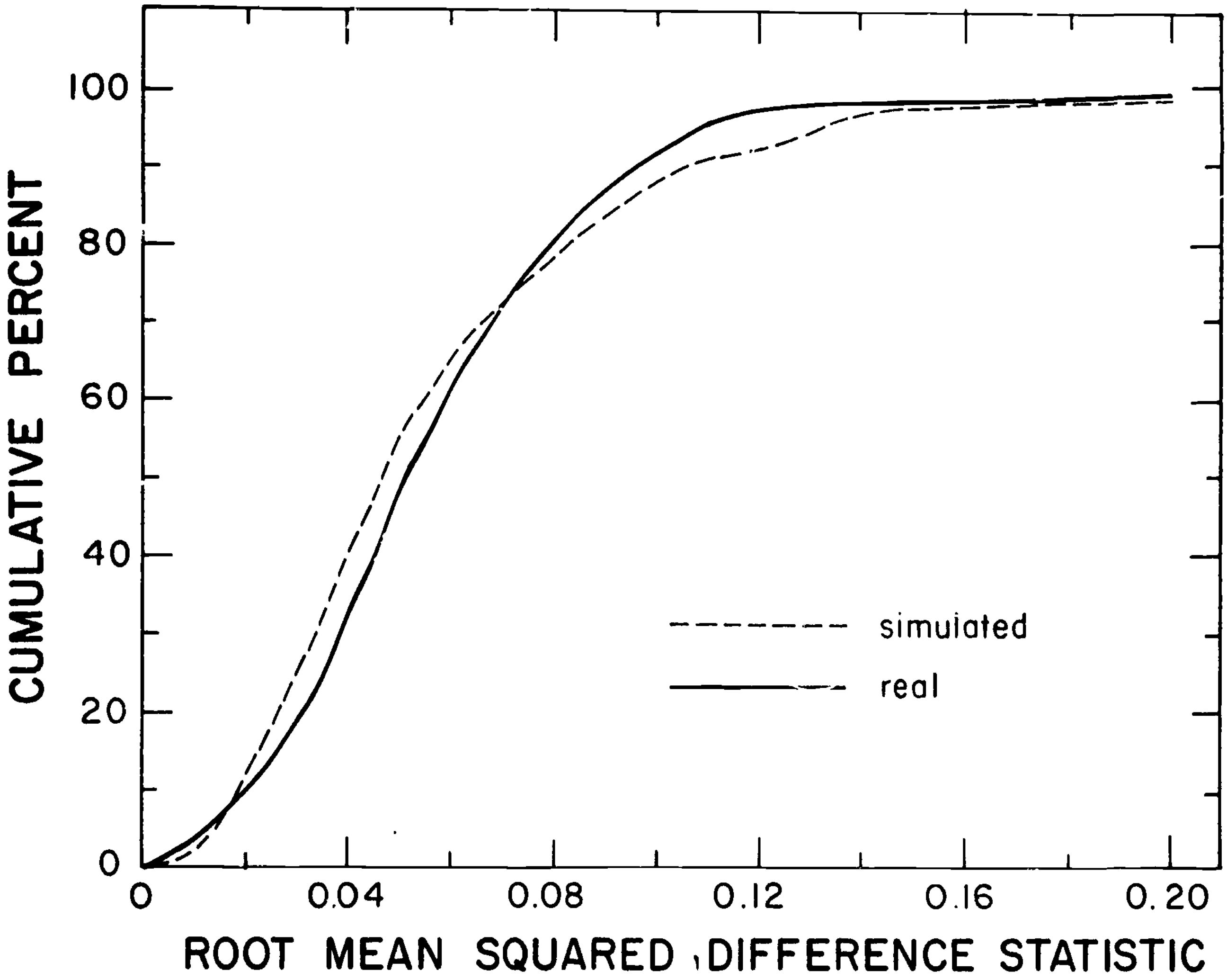
Figure 2. A comparison of the simulated and real sampling distributions of the Item Root Mean Squared Difference Statistics under the hypothesis of no bias.

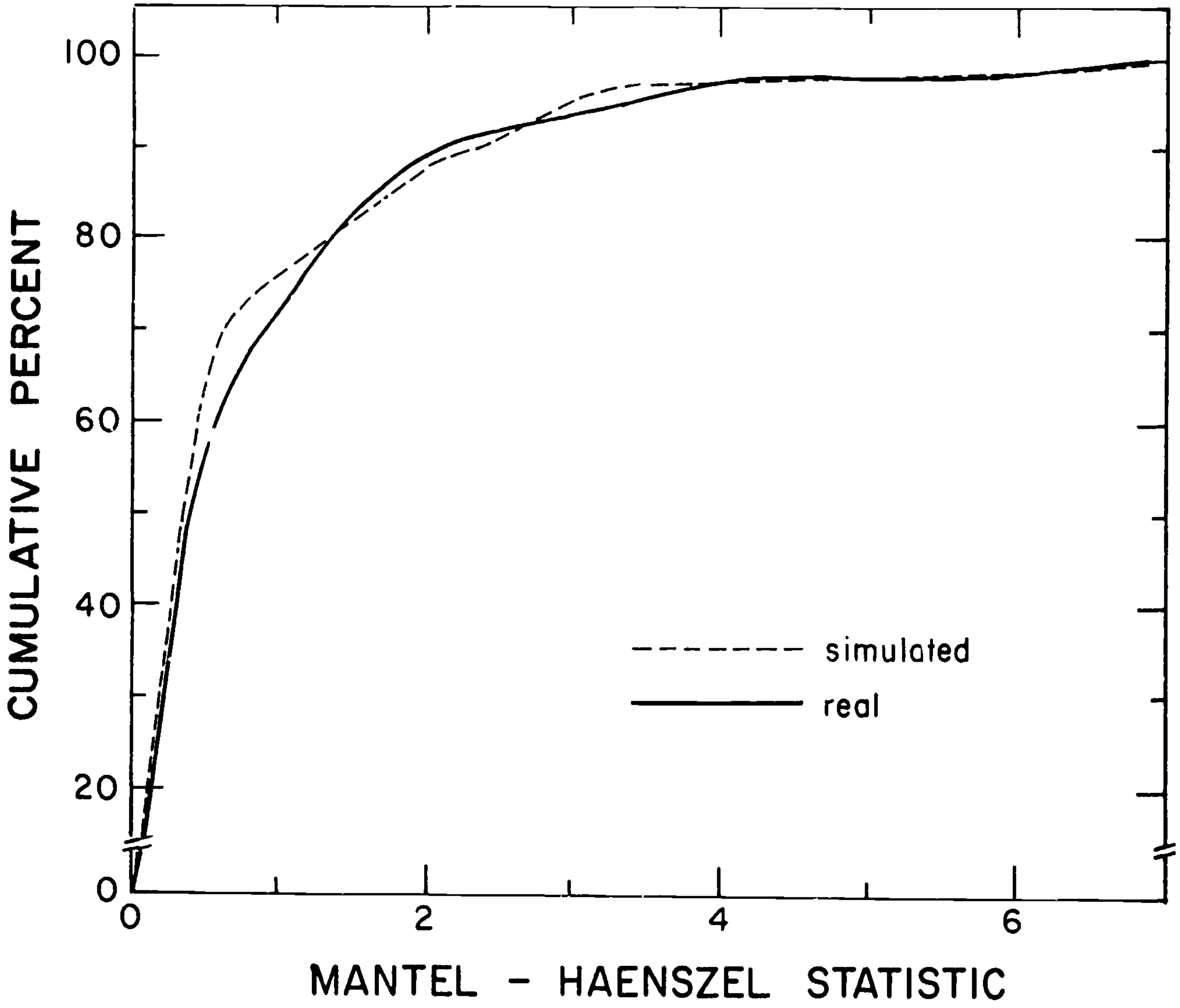
Figure 3. A comparison of the simulatod and real sampling distributions of the Item Mantel-Haenszel Statistics under the hypothesis of no bias.

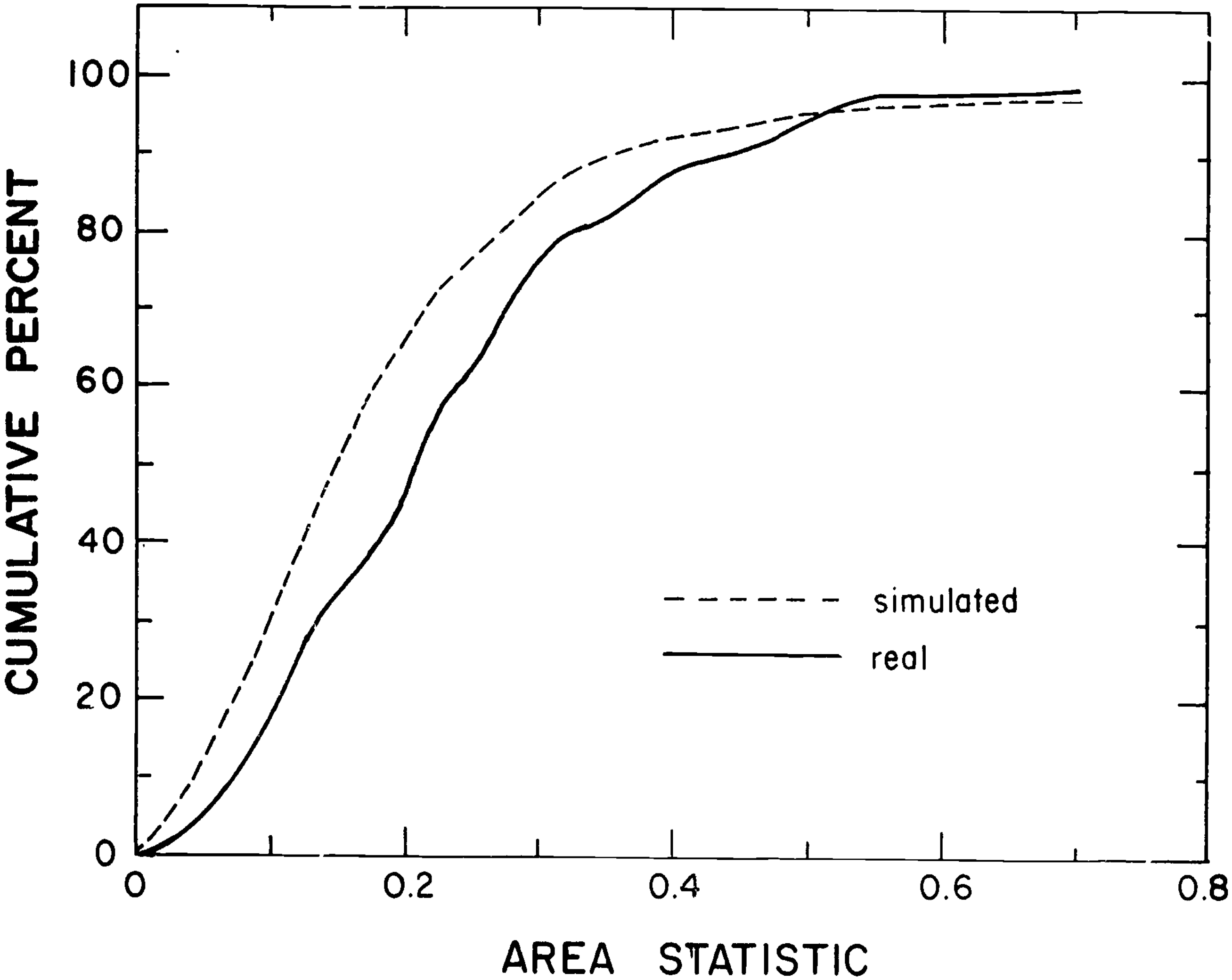
Figure 4. A comparison of the distribution of Item Area Statistics for the total male and female groups, and the smoothed distribution of the same statistic for the total simulated male and female groups under the hypothesis of no bias.

JANE.2.22


CUMULATIVE PERCENT
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
AREA STATISTIC
simulated
real


100
80
60
40
20
0
CUMULATIVE PERCENT
0
0.04
0.08
0.12
0.16
0.20
ROOT MEAN SQUARED DIFFERENCE STATISTIC
simulated
real


100
80
60
40
20
0
CUMULATIVE PERCENT
0
2
4
6
MANTEL - HAENSZEL STATISTIC
simulated
real


CUMULATIVE PERCENT
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
AREA STATISTIC
simulated
real