DOCUMENT RESUME

ED 287 870 TM 870 592

AUTHOR Rogers, H. Jane; Hambleton, Ronald K.
TITLE Evaluation of Computer Simulated Baseline Statistics for Use in Item Bias Studies.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
REPORT NO LPER-162
PUB DATE 29 Jun 87
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Simulation; *Cutting Scores; Grade 9; *Latent Trait Theory; Mathematical Models; Reading Tests; Sampling; Secondary Education; *Sex Bias; Sex Differences; Statistical Studies; *Test Bias; Test Items
IDENTIFIERS *Mantel Haenszel Procedure; *Root Mean Square (Statistics)

ABSTRACT

Though item bias statistics are widely recommended for use in test development and analysis, problems arise in their interpretation. This research evaluates logistic test models and computer simulation methods for providing a frame of reference for interpreting item bias statistics. Specifically, the intent was to produce simulated sampling distributions of item bias statistics under the no-bias hypothesis, for use in determining cut-off points to provide guidelines for interpreting item bias statistics obtained with actual test data. In this case, potential sex bias was studied in the item responses of 937 Cleveland ninth graders to 75 items from the 1985 Cleveland Reading Competency Test. Results supported the basic data simulation approach used in the study. Real and simulated distributions for three item bias statistics (area between characteristic curves, root mean squared differences between curves, and the Mantel-Haenszel statistic) when bias was not present were very similar. The minor differences found between the distributions had little effect on the interpretation of item bias statistics obtained with actual data. Seven steps for applying the method of computer-simulated baseline statistics in test development settings were outlined. (Author/LPG)

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

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 value 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.

JANE.2.1

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

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

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 are based upon item response models and related procedures and 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. But test developers often want to sort test items into ordered categories (e.g., "must be very carefully reviewed", "may need revision", "should be acceptable") and for this purpose, 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

---

[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.

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 purpose of the present research was to evaluate the value of logistic test models and computer simulation methods for generating sampling distributions of item bias statistics under the hypothesis of no item bias, for use in determining cut-off points to provide guidelines for interpreting item bias statistics.

This study was prompted by some earlier research by Hambleton, Rogers and Arrasmith (1986). These authors used baseline data for interpreting item bias statistics which were provided by two randomly equivalent majority samples, and two randomly equivalent minority samples, obtained from real data. This meant that while meaningful baseline results were available, the important comparisons between the majority and minority groups were carried out with sample sizes half the size 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 sample sizes are often not very large to begin with. Small sample item bias studies are especially problematic when IRT methods are used. The results from the Hambleton et al. study showed that logistic models could be used to provide simulated results to serve as a baseline for interpreting item bias statistics. But it was also clear 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 conducting an item bias investigation using two randomly equivalent samples drawn from the combined sample (Shepard, Camilli, & Williams, 1984; Wilson-Burt, Fitzmartin, & Skaggs, 1986). Since 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, and it is 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. Since group ability distributions can influence the quality of item bias statistics (see for example, Shepard et al., 1984; Wilson-Burt, Fitzmartin, & Skaggs, 1986), failure to incorporate this information in the analysis could reduce the usefulness of the distribution of item bias statistics obtained with 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 (see, for example, 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, our 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.

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, & 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, & Wardrop, 1981), the squared difference between the majority and minority item characteristic curves at fixed intervals (usually .01) is calculated. 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, and items associated with large-valued statistics are labeled as "potentially biased."

Mantel-Haenszel Method. The Mantel-Haenszel statistic 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 & Thayer, 1986). Unlike the other two methods, this method does not involve the application of 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 n items, there are 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) 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 described above is that the statistic has a known sampling distribution (chi-square with one degree of freedom) and so 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 responses of 937 Cleveland ninth grade students to 75 items on the 1985 Cleveland Reading Competency Test. There were 207 Whites and 730 Blacks; and 451 Males and 486 Females in the sample. Because of the very small number of Whites in the sample, only a sex bias study was carried out.

Generation of Simulated Examinee Item Responses

Basically, the approach was to simulate examinee item response data that reflected as closely as possible the actual examinee and item

data of interest but 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 responses for the 937 examinees was generated using the three-parameter logistic model (Hambleton & Rovinelli, 1973). In this way, the simulated item responses 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 response data were generated from a common set of three-parameter item characteristic curves obtained from the analysis of the total set of test data. 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 were 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 since in all other respects the datasets were equivalent, assuming 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 & Rogers, in press; Hambleton & Swaminathan, 1985).

Procedure

With the actual and simulated test data in hand, three sets of analyses were carried out: 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) 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 using simulated rather than real sampling distributions in setting cut-off scores. This analysis involved 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 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. Here, the purpose was to highlight how the method can work in practice.

The specific steps in our 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 carried out. The distribution of these item bias statistics (no bias present) provided a basis for evaluating the distribution obtained from the simulated test data. Next, the simulated test data was 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 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. M_1 and F_1, and M_2 and F_2 comparisons were preferred to the M_1 and M_2, 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 carried out. 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. The Mantel-Haenszel statistics were calculated using the item response data provided at step 1.

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

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

Real Data

a. the combined distribution of M_1 vs M_2 and F_1 vs F_2 item bias statistics. (This distribution served as the baseline for interpreting the real item bias statistics obtained from the M_1 vs F_1 and M_2 vs F_2 comparisons.)

b. the distributions of the M_1 vs F_1, and the M_2 vs 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 M_1 vs F_1 and M_2 vs F_2 item bias statistics. (This distribution served as the alternate baseline for interpreting the real item bias statistics obtained from the M_1 vs F_1, and M_2 vs F_2 groups.) This distribution was compared to 4(a) obtained above to assess the viability of the computer-generated sampling distributions of item bias statistics.

JANE.2.9

5. The distributions obtained in step 4 (except for the real M vs F comparison) were smoothed by the method of "weighted rolling averages" (Kendall & Stuart, 1968) to remove some of the minor irregularities in the distributions.
6. The cut-off score corresponding to the .05 level of 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 response 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 absolute 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 best-fitting model and the actual test data.

Comparison of the Real and Simulated Null Distributions

Tables 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. Figures 1, 2, and 3 highlight the same

---

Insert Tables 1, 2, and 3 and Figures 1, 2, and 3 about here

---

information in graphical form. The results are 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 on the determination of cut-off values were small.

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 different cut-off scores derived from the two distributions. Table 4 provides the .05 cut-off score for the real and simulated distributions for each item bias statistic under the hypothesis of no bias. These

---

Insert Table 4 and Figures 4, 5, and 6 about here

---

cut-off scores were then applied to the M_1 vs F_1 and M_2 vs F_2 real item bias data. The smoothed simulated distribution without bias and the smoothed real distribution of item bias statistics for the M_1 vs F_1 item bias study are shown in Figures 4, 5, and 6.

Table 4 shows that there were differences in the determination of cut-off scores 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 .05 cut-off score 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 better 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 our computer simulation. In practice, a test developer would carry out the item bias study with the full set of available data. Figures 7, 8, and 9 highlight the results of the item bias investigation using the full Male and Female samples, and using smoothed computer-simulated distributions of item bias statistics without any bias, to provide baseline data for interpreting the results. Using the .05 level of significance, the numbers of items in need of careful review were obtained. The numbers varied depending on the choice of item bias statistic: eight items with the Area method, six items with the Root Mean Squared Difference method, and 20 items with the Mantel-Haenszel method.

Insert Figures 7, 8, and 9 about here

Conclusions

The results of this study reported in Tables 1 to 3 and 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 user to simulate samples closely resembling the original data but under conditions of no bias. Though the results in these tables and figures do not provide much 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, since 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 allowed us to set more realistic cutoff values for the bias statistics. Certainly, nothing was lost with the procedure and there may be circumstances in practice where there is considerable merit to the use of simulated distributions of item 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 using random samples of the real data, resulting in the flagging of fewer items as biased. Given that our groups were Males and Females, and 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 (see Table 4) highlights the problem of using IRT methods in small samples. This leads us to caution against using any firm cut-off score for the bias statistics. We recommend 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; however, we would still recommend that precise cut-off scores not be used.

The results for the Mantel-Haenszel statistic suggest that while 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 which closely resemble the real data. Hence the method proposed here of simulating data to obtain baseline values may not be useful for bias statistics which 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. Choose a new, better fitting model, otherwise, and repeat this step. 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 responses using the logistic model of choice in step 1 (see, for example, Hambleton & inelli, 1973).

JANE.2.15

3. Split the simulated examinee item responses into the majority and minority groups of interest and re-estimate the item parameters, treating ability scores obtained at step 1 as fixed. (Fixing the ability scores serves two purposes: item parameter estimation time is reduced substantially, and 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 using the cut-off values obtained from the simulated test data.

Test developers who carry out the seven steps above should be in a position to interpret their item bias statistics in a meaningful way. Our view remains, however, that while simulated sampling distributions can be very useful when interpreting actual item bias statistics, because of the instability of determining cut-off scores as well as the arbitrariness of the choice of cut-off scores, judgment must still be used in making sensible use of item bias statistics.

References

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

Hambleton, R.K., & 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., & Arrasmith, D. (1986). 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., & Rovinelli, R.J. (1973). A Fortan IV program for generating examinee response data for logistic test models. Behavioral Science, 18, 74.

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

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

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

Linn, R.L., Levine, M.V., Hastings, C.N., & 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., & 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., & 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., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L., Camilli, G., & 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., & Skaggs, G. (1986). Baseline strategies in evaluating IRT item bias indices. Paper presented at the annual meeting of AERA, San Francisco.

JANE.2.17

Table 1

Distribution of the Item Area Statistics Under the Hypothesis of No Bias

Interval (Mid-Point)	Real Data Cum%	Simulated Data Cum%	Cum% Difference
.015	1.0	0.3	0.7
.045	3.2	2.9	0.3
.075	6.6	8.1	1.5
.105	11.2	15.4	4.2
.135	17.2	23.8	6.6
.165	25.2	32.9	7.7
.195	34.4	41.4	7.0
.225	44.1	49.3	5.2
.255	53.4	57.3	3.9
.285	61.1	64.0	2.9
.315	67.5	69.2	1.7
.345	73.0	73.9	0.9
.375	77.8	77.7	0.1
.405	81.9	80.5	1.4
.435	85.7	83.4	2.3
.465	89.4	86.3	3.1
.495	92.5	88.8	3.7
.525	94.9	91.0	3.9
.555	96.6	92.6	4.0
.585	97.6	93.4	4.2
.615	98.0	94.2	3.8
.645	98.0	95.0	3.0
.675	98.1	96.0	2.1
.705	98.1	96.8	1.3
.735	98.3	97.4	0.9
.765	98.5	97.8	0.7
.795	98.8	98.3	0.5
.825	99.3	99.0	0.3
.855	100.0	99.6	0.4

JANE.1.1

Table 2

Distribution of the Item Root Mean Squared Difference Statistics Under the Hypothesis of No Bias

Interval (Mid-Point)	Real Data Cum%	Simulated Data Cum%	Cum% Difference
.0025	0.6	0.6	0.0
.0075	2.2	0.8	1.4
.0125	4.6	3.6	1.0
.0175	7.8	8.4	0.6
.0225	11.3	14.4	3.1
.0275	15.3	20.5	5.2
.0325	20.6	27.5	6.9
.0375	27.4	35.2	7.8
.0425	35.6	43.1	7.5
.0475	44.1	51.1	7.0
.0525	51.4	57.5	6.1
.0575	57.8	62.5	4.7
.0625	63.4	67.0	3.6
.0675	68.7	70.8	2.1
.0725	73.7	73.8	0.1
.0775	78.4	76.8	1.6
.0825	82.6	79.8	2.8
.0875	85.6	82.4	3.2
.0925	88.1	84.7	3.4
.0975	90.7	87.2	3.5
.1025	92.8	89.3	3.5
.1075	94.7	90.7	4.0
.1125	96.5	91.8	4.7
.1175	97.5	92.4	5.1
.1225	97.9	93.0	4.9
.1275	98.1	94.2	3.9
.1325	98.2	95.6	2.6
.1375	98.3	96.9	1.4
.1425	98.5	97.7	0.8
.1475	98.8	98.1	0.7
.1525	99.0	98.1	0.9
.1575	99.2	98.1	1.1
.1625	99.3	98.3	1.0
.1675	99.3	98.5	0.8
.1725	99.4	98.8	0.6
.1775	99.6	99.0	0.6
.1825	100.0	99.2	0.8

JANE.1.2

Table 3

Distribution of the Item Mantel-Haenszel Statistics
Under the Hypothesis of No Bias

Interval (Mid-Point)	Real Data Cum%	Simulated Data Cum%	Cum% Difference
.1	20.7	25.2	4.5
.3	42.6	49.1	6.5
.5	58.1	65.3	7.2
.7	65.9	72.6	6.7
.9	69.8	74.7	4.9
1.1	74.6	77.6	3.0
1.3	79.3	80.1	0.8
1.5	83.3	82.7	0.6
1.7	86.4	85.5	0.9
1.9	88.6	87.9	0.7
2.1	90.1	89.2	0.9
2.3	91.1	90.0	1.1
2.5	92.1	91.4	0.7
2.7	93.2	93.2	0.0
2.9	93.7	94.9	1.2
3.1	94.2	96.1	1.9
3.3	94.9	96.9	2.0
3.5	95.7	97.1	1.4
3.7	96.5	97.2	0.7
3.9	97.4	97.3	0.1
4.1	97.9	97.3	0.6
4.3	98.0	97.4	0.6
4.5	98.0	97.5	0.5
4.7	98.1	97.6	0.5
4.9	98.1	97.9	0.2
5.1	98.1	98.1	0.0
5.3	98.2	98.4	0.2
5.5	98.3	98.5	0.2
5.7	98.5	98.7	0.2
5.9	98.7	98.7	0.0
6.1	98.7	98.7	0.0
6.3	98.8	98.7	0.1
6.5	99.1	98.8	0.3
6.7	99.5	99.1	0.4
6.9	100.0	99.5	0.5

Table 4

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

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

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

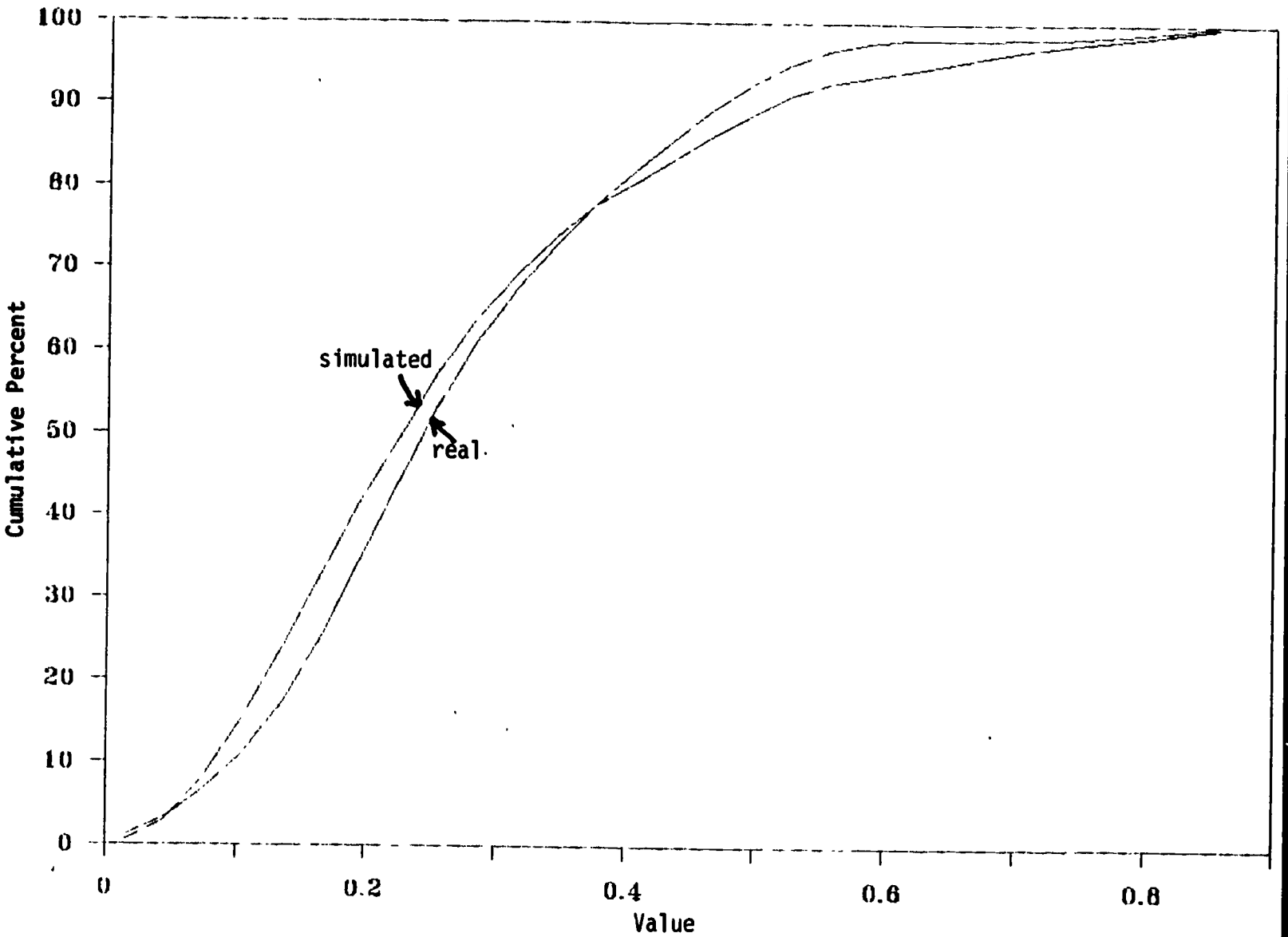

Figure 2. A comparison of the simulated and real sampling distribution of the Item Root Mean Squared Difference Statistics under the hypothesis of no bias.

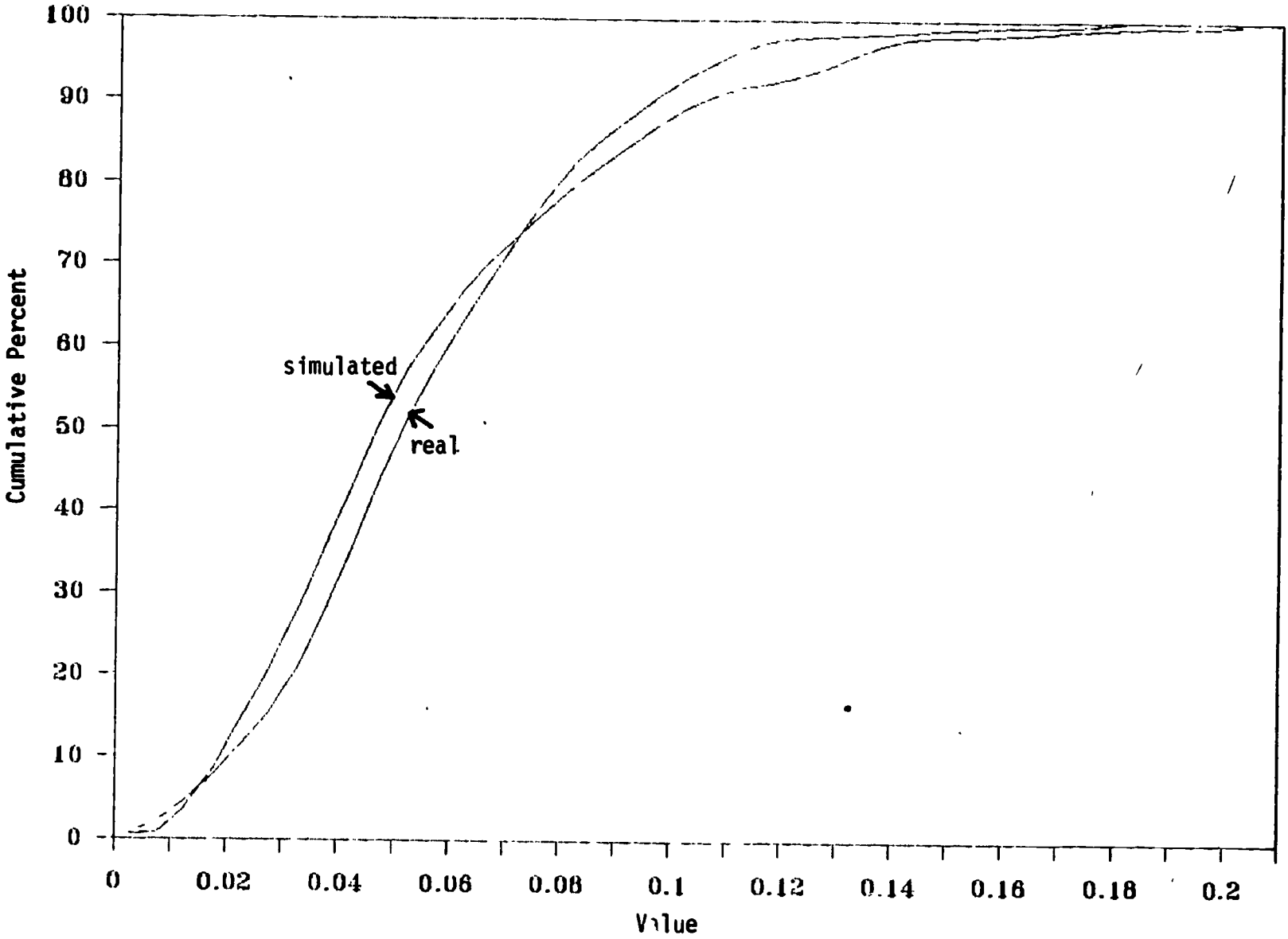

Figure 3. A comparison of the simulated and real sampling distributions of the Item Mantel-Haenszel Statistics under the hypothesis of no bias.

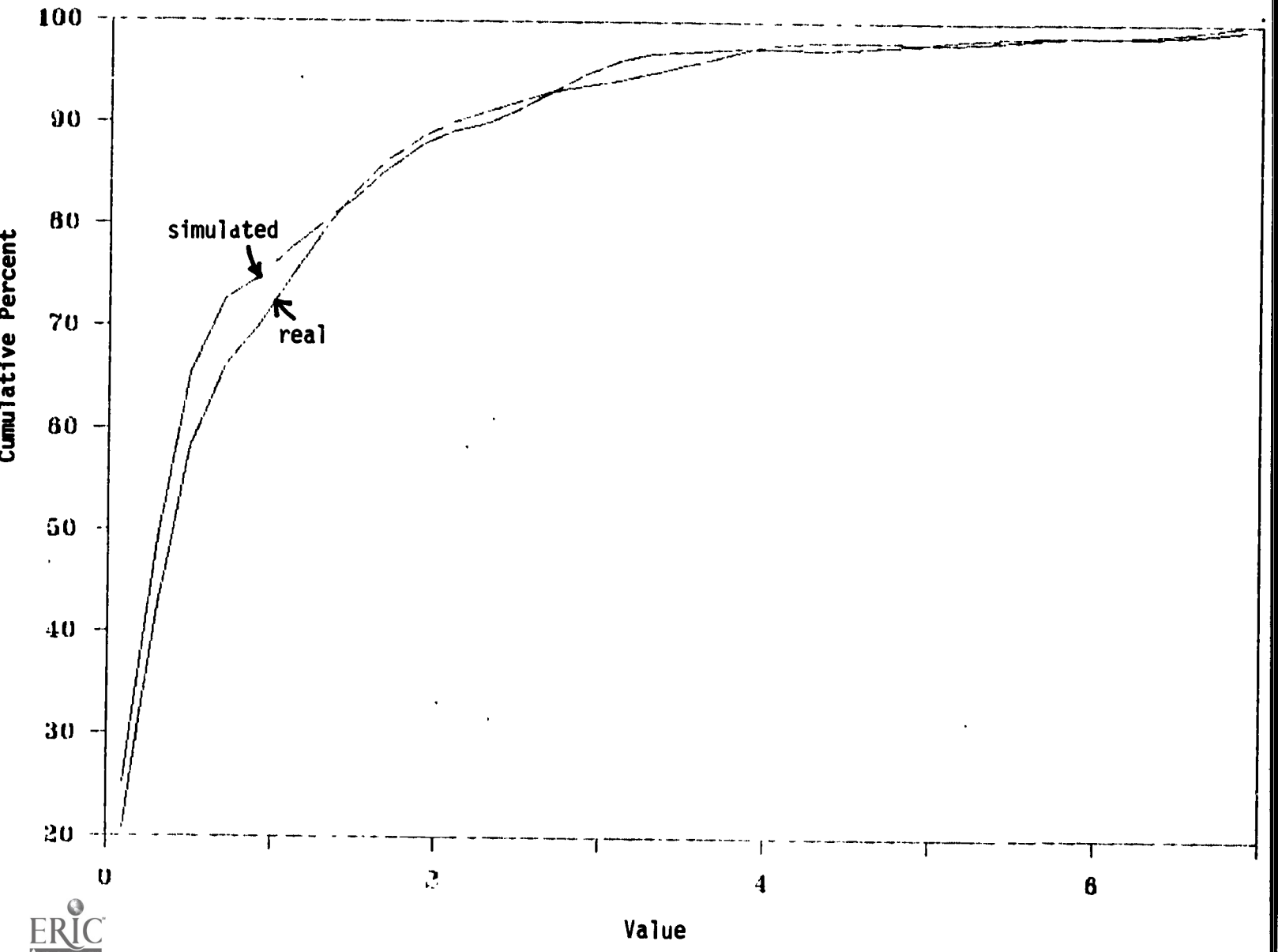

Figure 4. A comparison of the distribution of Item Area Statistics for the male and female groups with the smoothed distribution of the same statistic for the simulated male and female groups under the hypothesis of no bias.

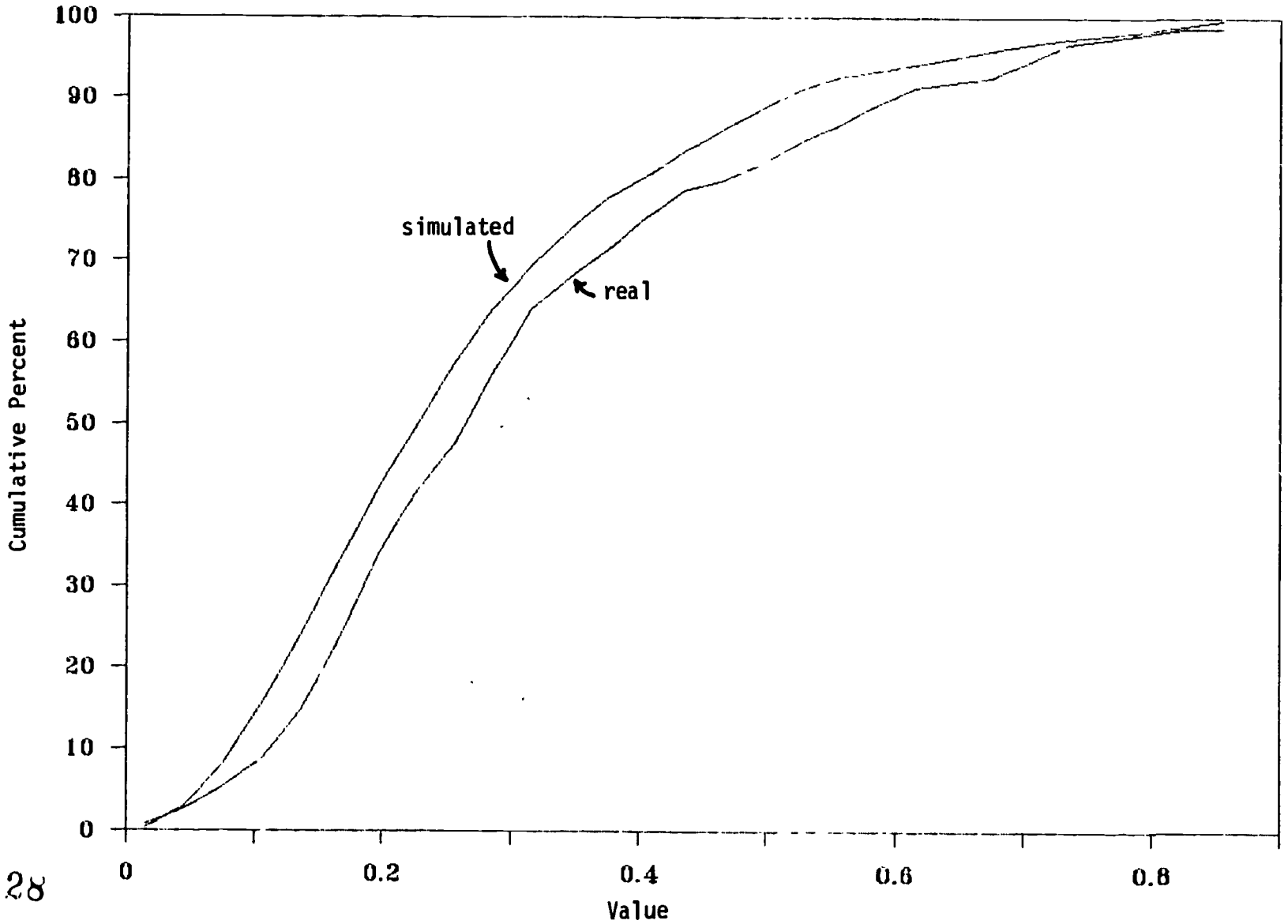

Figure 5. A comparison of the distribution of Item Root Mean Squared Difference Statistics for the male and female groups with the smoothed distribution of the same statistic for the simulated male and female groups under the hypothesis of no bias.

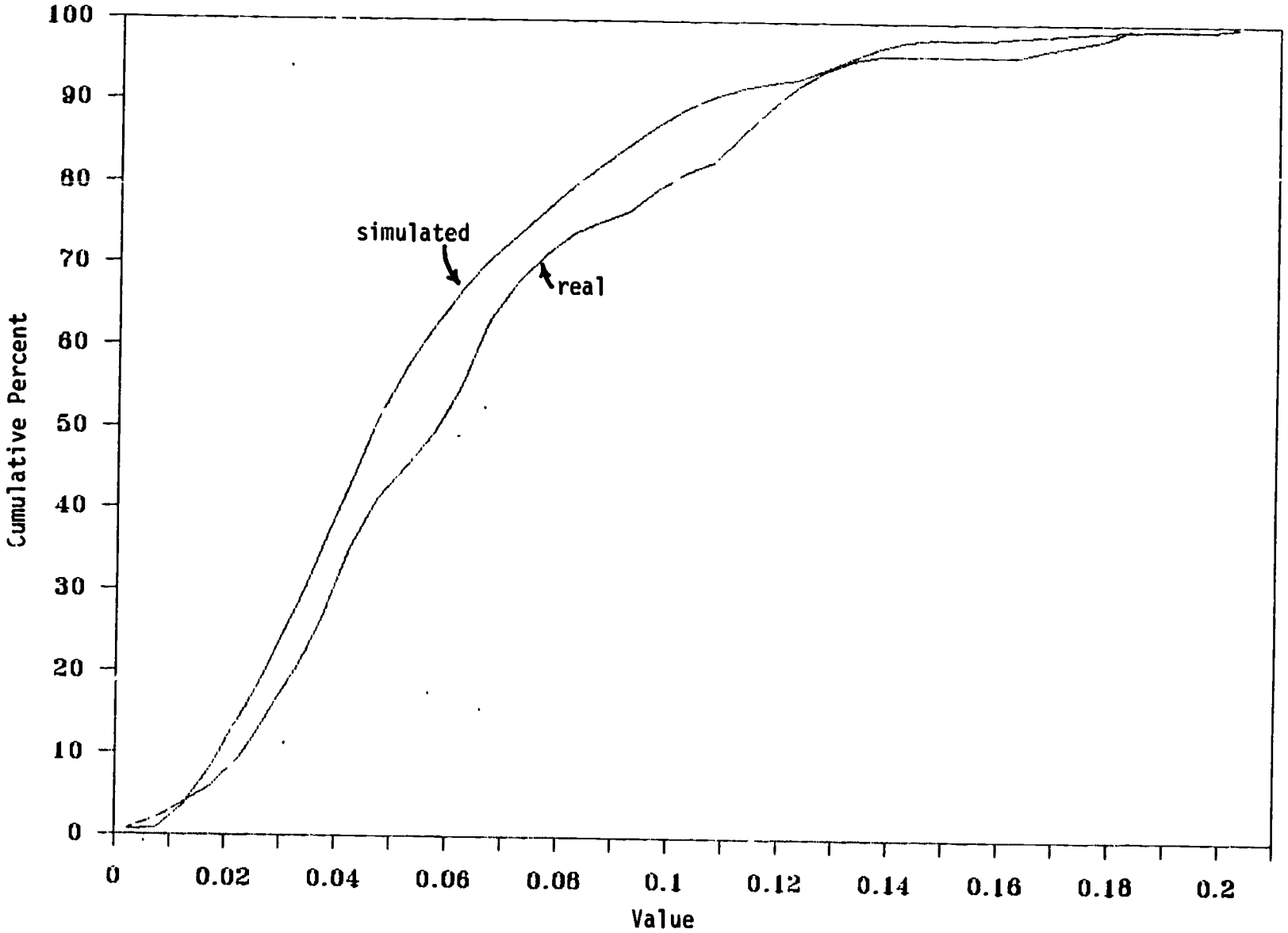

Figure 6. A comparison of the distribution of Mantel-Haenszel Statistics for the male and female groups with the smoothed distribution of the same statistic for the simulated male and female groups under the hypothesis of no bias.

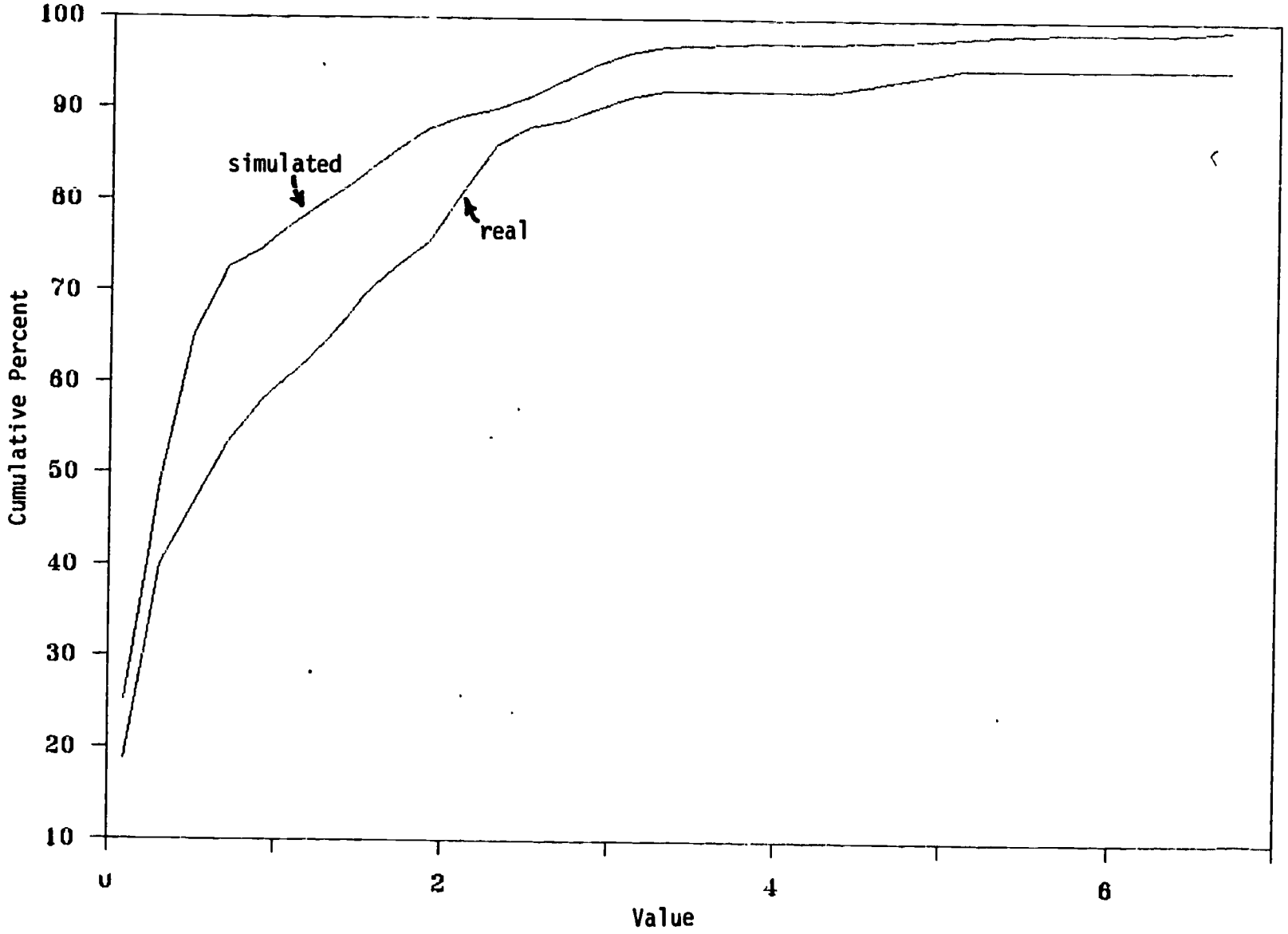

Figure 7. A comparison of the distribution of Item Area Statistics for the total sample male and female groups with the smoothed distribution of the same statistic for the total sample simulated male and female groups under the hypothesis of no bias.

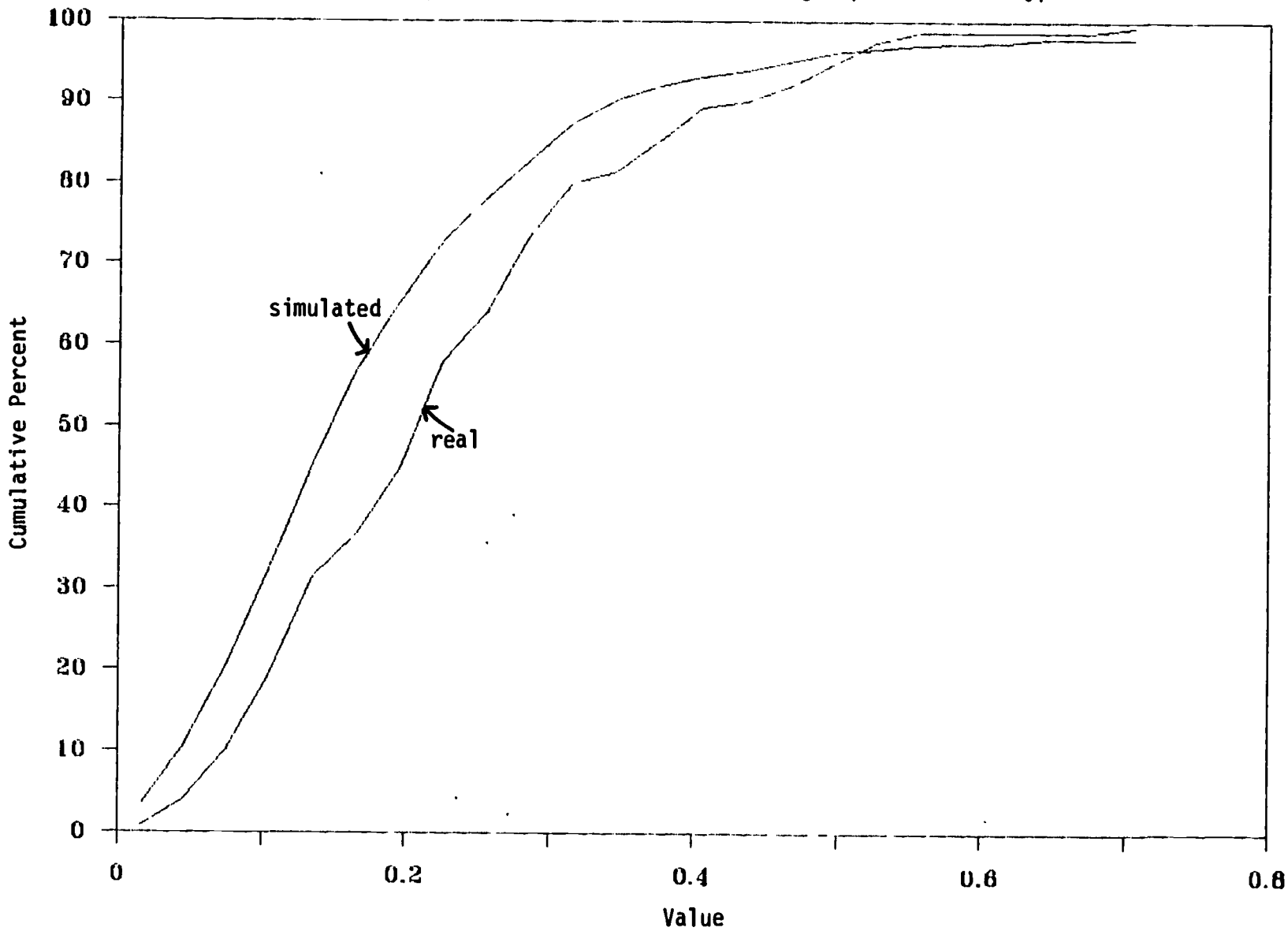

Figure 8. A comparison of the distribution of Item Root Mean Squared Differences for the total sample male and female groups with the smoothed distribution of the same statistic for the total sample simulated male and female groups under the hypothesis of no bias.

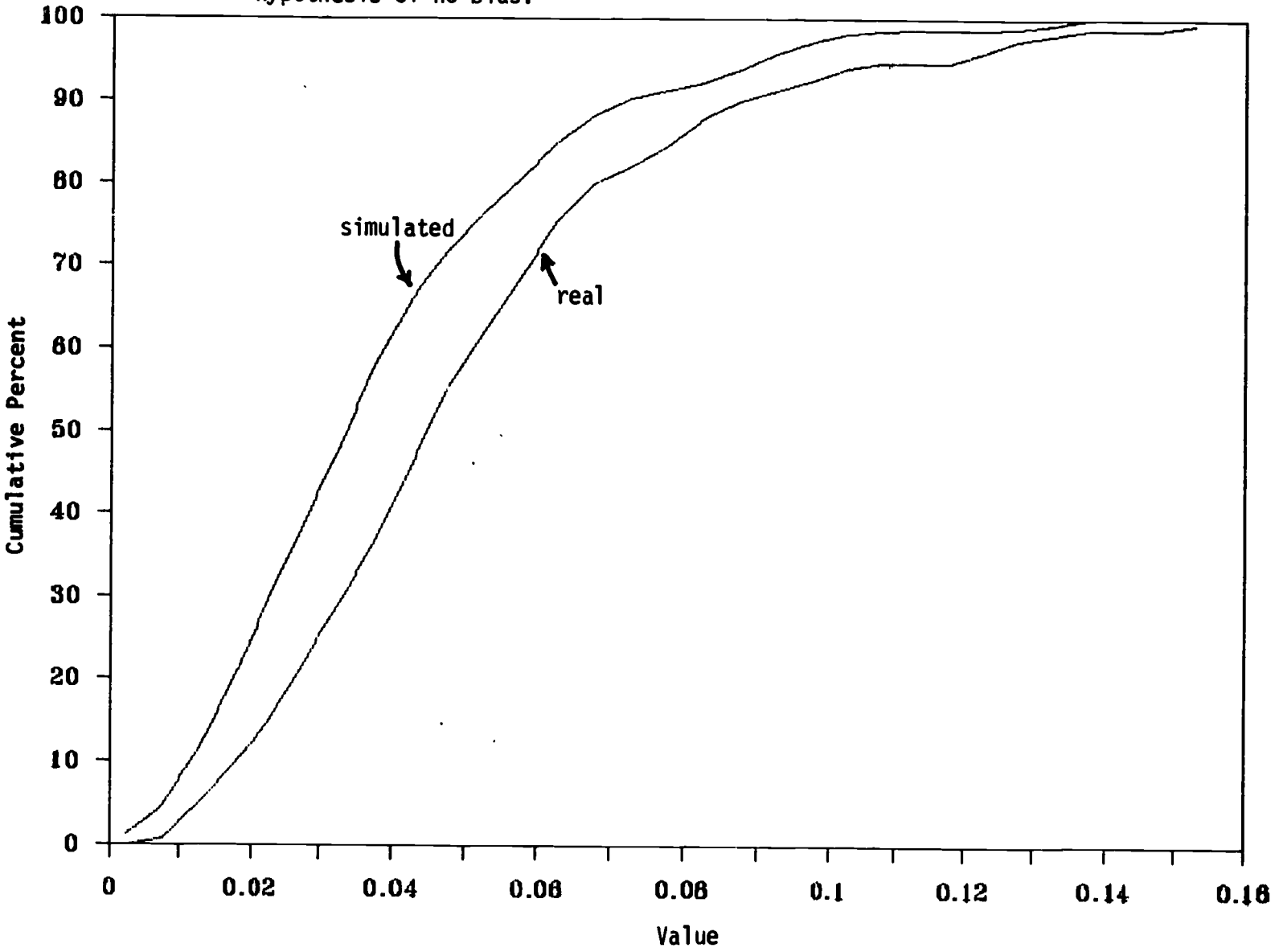

Figure 9. A comparison of the distribution of Mantel-Haenszel Statistics for the total sample male and female groups with the smoothed distribution of the same statistic for the total sample simulated male and female groups under the hypothesis of no bias.