ACT Research Report Series 92-1

Performance of the Mantel-Haenszel Statistic and the Standardized Difference in Proportions Correct When Population Ability Distributions are Incongruent

Judith A. Spray
Timothy R. Miller

July 1992

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

Performance of the Mantel-Haenszel Statistic and the Standardized Difference in Proportions Correct When Population Ability Distributions are Incongruent

Judith A. Spray and Timothy R. Miller

American College Testing

Abstract

A popular method of analyzing test items for differential item functioning (DIF) is to compute a statistic that conditions samples of examinees from different populations on an estimate of ability. This conditioning or matching by ability is intended to produce an appropriate statistic that is sensitive to true differences in item functioning, provided the ability estimate accurately reflects a comparable level of the true ability for these populations. If the observed or number-correct score is used as a conditioning or grouping variable, a problem exists whenever examinees from two different populations are matched on the same level of the observed test score, but actually have quite different levels of the unobserved ability. This occurs whenever the distributions of true abilities for the populations of interest are incongruent or non-overlapping. This situation was investigated in a series of computer simulations. The results indicated that the magnitude of the problem, in terms of being able to detect true DIF with moderate sample sizes when ability distributions are incongruent, may not be that serious for tests which are, on average, free from DIF.

Performance of the Mantel-Haenszel Statistic and the Standardized Difference in Proportions Correct When Population Ability Distributions Are Incongruent

Two statistics that are used to indicate differential item functioning (DIF) between two populations of examinees are the Mantel-Haenszel common-odds ratio (MH) (or equivalently the Mantel-Haenszel negative log-odds ratio) (Holland & Thayer, 1986) and the standardized difference (STD) in proportions correct (Dorans & Kulick, 1986). Both statistics condition on some ability measure, usually the observed score of the test containing the items undergoing the DIF analysis. Conditioning on the observed test score in order to evaluate population differences in item proportion correct would appear to be appropriate provided the matching observed test score accurately reflects a comparable level of the measured trait for the populations of interest. However, problems arise whenever identical values of the observed test score, *X*, represent different levels of ability across groups. This can occur when the conditional distributions of ability given observed score are different for the comparison groups used in the DIF analysis.

Zwick (1990) has discussed the implications of this problem within a theoretical context. The purpose of the current paper is to present a more applied analysis of this problem and to attempt to determine how severe the situation must be before a DIF analysis that employs the MH or STD statistic leads to erroneous conclusions.

Definitions of the DIF Statistics

The definitions of the estimator of the standardized difference in proportions correct (STD) and the Mantel-Haenszel common-odds ratio estimator (MH) are given as follows.

If the two populations of examinees are labeled as a focal group (*F*) and a base group (*B*), and *s* indexes each observed score category of a *k*-item test, or $s = 0, 1, ..., k$, then

N_{F_s} - the number of examinees in the *F* group at score *s*,

N_{B_s} - the number of examinees in the *B* group at score *s*,

N_s - the number of examinees in *F* and *B* at score *s*,

$G_{F_s} - N_{F_s} / \sum_{s-0}^{k} N_{F_s}$, the relative frequency of *F* at *s*,

$G_{B_s} - N_{B_s} / \sum_{s-0}^{k} N_{B_s}$, the relative frequency of *B* at *s*, and

$G_s - N_s / \sum_{s-0}^{k} N_s$, the total relative frequency of *F* and *B* at *s*.

If R_{F_s} and R_{B_s} are the numbers of examinees (i.e., absolute frequency), in *F* and *B* respectively, at *s* who answer the item of interest correctly, then the proportion-correct values for each group at *s* are given by $P_{F_s} - R_{F_s} / N_{F_s}$, and $P_{B_s} - R_{B_s} / N_{B_s}$.

The STD Statistic

The standardized difference in proportions correct is defined as

$$STD - \sum_{s-0}^{k} (P_{F_s} - P_{B_s})\, G_{F_s}, \qquad (1)$$

where the signed difference, $P_{F_s} - P_{B_s}$, is weighted by the relative frequency of F. The statistic is defined on the proportion-correct scale and indicates, on average, how members of F differed from comparable members of B. Negative values of STD indicate that an item favors B, while positive values indicate that an item favors F. Values of the STD statistic near zero indicate no DIF.

The MH Statistic

If W_{F_s} and W_{B_s} are the absolute frequencies of incorrect responses to this item in F and B respectively at s, and N_s is the total number of responses at s, then the Mantel-Haenszel common-odds ratio estimator is

$$MH = \frac{\sum_{s=0}^{k} R_{B_s} W_{F_s} / N_s}{\sum_{s=0}^{k} R_{F_s} W_{B_s} / N_s} . \qquad (2)$$

If Q_{F_s} and Q_{B_s} are defined as $(1 - P_{F_s})$ and $(1 - P_{B_s})$ respectively, then this statistic could also be written as

$$MH = \frac{\sum_{s=0}^{k} P_{B_s} Q_{F_s} \frac{G_{B_s} \cdot G_{F_s}}{G_s}}{\sum_{s=0}^{k} P_{F_s} Q_{B_s} \frac{G_{B_s} \cdot G_{F_s}}{G_s}} . \qquad (3)$$

The MH statistic can be interpreted as an estimate of the common odds-ratio. It indicates, on average, how much more (or less) likely it is that a member of B answered the item correctly than did a comparable member of F. The MH statistic has a value at or near 1.0 if there is no DIF between B and F. If the item favors B, MH is greater than

1.0; if the item favors F, MH is less than 1.0. Frequently, the odds ratio is transformed by using some function of the negative of the natural log of the ratio.

Population DIF Indices

The DIF statistics given by Equations 1 and 2 are defined in terms of the observed test score. As mentioned previously, the examinees from each group ideally should be matched on their latent abilities or true scores. For computer simulation work, it is possible to define some measure of population DIF in terms of the latent ability or true score. This value then becomes the parameter of interest in estimation because it represents the value of the indices when true ability matching or conditioning has occurred. The DIF statistics can be compared to these population DIF indices, which then serve as a reference for valid DIF identification.

The usual assumption concerning the latent ability or true score can be made, namely that the latent ability, θ, is a continuous random variable with known density functions. If these arbitrary density functions of θ are denoted by $g_F(\theta)$ and $g_B(\theta)$, then the combined group density can be represented by

$$g^*(\theta) = \alpha g_F(\theta) + (1-\alpha)g_B(\theta) ,$$

where a mixing proportion, α, is defined as $0 \leq \alpha \leq 1$. The mixing proportion is usually taken to equal the relative proportion of examinees who appear in F (either sampled or in the F population).

The definitions of each population DIF index are facilitated by replacing the proportions correct and incorrect at each score category (i.e., P_{B_S}, Q_{F_S}, P_{F_S} and Q_{B_S}) with probability functions of the latent ability variable, θ. In the context of the present paper, it was assumed that the success probabilities, $P_B(\theta)$ and $P_F(\theta)$, were given by the unidimensional three-parameter logistic item response function with known item parameters for each group and for each item, or in general by

$$P(\theta) = c + \frac{(1-c)}{1 + e^{-1.7a(\theta-b)}} \; . \qquad (4)$$

A population value of STD, μ_{STD}, was defined as the expected difference between the proportions correct, relative to (or weighted by) $g_F(\theta)$ as the standardizing distribution (Kendall, Stuart, & Ord, 1987, p. 46), or

$$\mu_{STD} = \int_{-\infty}^{\infty} [P_F(\theta) - P_B(\theta)] \, g_F(\theta) \, d\theta \; . \qquad (5)$$

The population value of the common-odds ratio, ψ, was defined to be the latent variable-equivalent to Equation 3, or

$$\psi = \frac{\int_{-\infty}^{\infty} P_B(\theta) Q_F(\theta) \frac{g_B(\theta) g_F(\theta)}{g^*(\theta)} \, d\theta}{\int_{-\infty}^{\infty} P_F(\theta) Q_B(\theta) \frac{g_B(\theta) g_F(\theta)}{g^*(\theta)} \, d\theta} \; . \qquad (6)$$

Defining Equation 6 as the population value of the common-odds ratio is not without some interpretative difficulties. Greenland (1982) pointed out that, although there are several interpretations of an odds ratio when the ratio is not assumed to be homogeneous in the population (i.e., the odds ratio is not constant across different values of θ), the weights, $(G_{B_s} G_{F_s})/G^*_s$, used in the Mantel-Haenszel estimator have no logical interpretation in the population. However, within the context of the present study, it was more important to compare the effects of conditioning on the observed score as opposed to the true latent variable, θ, than to defend one population interpretation over another. And because different definitions of the population odds ratio can result in quite discrepant values for the odds ratio parameter (Greenland, 1982), the definition given in Equation 6 was chosen so that any confounding of results which could be attributed to an inconsistent choice of the population odds ratio (i.e., inconsistent with the MH statistic) would be eliminated.

Prior Ability Distributions

Previously, it was stated that if examinees from different populations have been matched on observed test scores, they might not be matched on latent abilities. This occurs whenever the conditional distributions of true score given observed score are different for the two groups.

Zwick (1990) showed that if the test reliabilities for both groups were less than 1.0, and if the means of the ability or true-score distributions for each group were not equal so that the ability distributions were incongruent, then the conditional distributions of true score given observed score would not be identical but would result in conditional distributions that were described as being *stochastically ordered*. Under certain circumstances, this could produce results that would lead to the MH DIF statistic erroneously favoring the group with the higher ability. Regardless of the order of the conditional error distributions of observed score given ability, or $f(X|\theta)$, if such distributions exist for both groups, then different distributions of ability, $g(\theta)$, will yield different conditional distributions of ability given observed score, $j(\theta|X)$, due to Bayes theorem.

Degree of Distributional Incongruence

A measure of the degree with which the two distributions of θ, $g_F(\theta)$ and $g_B(\theta)$, are incongruent is the percentage of overlap of the areas under the density functions. This measure allows for an infinite number of combinations of distributions to be mapped to a simple scalar between 0.0 (signifying no overlap or total incongruence) and 1.0 (complete overlap, or total congruence), and is defined by

$$\text{OVERLAP} - \int_{-\infty}^{\infty} \text{MIN}[g_B(\theta),g_F(\theta)]\ d\theta\ . \tag{7}$$

Method

The present study utilized computer simulation methods in order to manipulate the primary condition of interest, the degree of incongruence or overlap between the distributions of ability of two populations of examinees (*B* and *F*). In order to make the results more generalizable to real testing situations, item responses taken from previous administrations of a 40-item ACT Assessment Mathematics Usage Test were fit using a three-parameter logistic model that assumed a unidimensional examinee trait or ability. Two comparison samples of 2000 Caucasian and 2000 African-American examinees were used to obtain separate *B* and *F* group item parameter estimates. Marginal maximum likelihood procedures, which assumed standard normal prior ability distributions, were used on each of the two samples via the computer program, *PC-BILOG 3* (Mislevy & Bock, 1989).

Because the groups were thought to be nonequivalent, the item parameters from the *F* group (a_F, b_F, and c_F) were rescaled (a^*_F, b^*_F, and c^*_F) to the *B* group parameters using the family of linear transformations,

$$a^*_F = \frac{a_F}{A}\,, \qquad b^*_F = b_F{\cdot}A + B\,, \qquad c^*_F = c_F\,,$$

where

$$A = \frac{\mathrm{SD}(b_B)}{\mathrm{SD}(b_F)}\,, \text{ and } B = (\overline{b_B}) - A\cdot(\overline{b_F}).$$

"Real" DIF between the two groups on any of the 40 items was thus somewhat reflected in the item parameter estimates.[1] As far as goodness-of-fit was concerned, no statistical procedure was used to assess the degree of model fit or misfit. Prior experience has shown that the unidimensional three-parameter logistic model fits these types of mathematics items on samples of 2000 at least well enough to yield item parameters that can subsequently produce observed score distributions that are very close to those obtained from national administrations of the tests. Therefore, these parameters estimates were used as known item parameters in all of the subsequent computer simulations.

The *B* ability distribution, $g_B(\theta)$, was always assumed to be standard normal. Therefore, only $g_F(\theta)$ varied throughout the simulations, and the measure of incongruence between the two ability distributions was the proportion of their overlap (in area). The F ability distribution, $g_F(\theta)$, was normally distributed with variance fixed at either 1.0 or .5. The Focal group mean was varied such that $\mu_F(\theta) \le \mu_B(\theta)$.

The known item parameters were used to describe the success probabilities, $P_F(\theta)$ and $P_B(\theta)$, with the item response function given by Equation 4. Once $g_F(\theta)$ and $g_B(0)$ were specified and a value of θ from either F or B had been sampled, 40 item responses were generated in the usual way by comparing either $P_F(\theta)$ or $P_B(\theta)$ to a pseudorandomly generated uniform deviate between 0 and 1. Statistics were then computed as functions of the item responses from either Equation 1 and 2, and these values were compared to μ_{STD} and ψ from Equations 5 and 6, respectively. Actually, the negative of the natural log of Equation 2 and Equation 6 was computed. Sampling variability was achieved by replicating each simulation 100 times and by drawing samples of 500 values each of θ from $g_F(\theta)$ and $g_B(\theta)$.

Two methods were used to assess the fidelity of either DIF statistic in the identification of an item's true DIF status. One was to compute the bias, standard error and root mean square error relative to the population DIF value over replications. These values also could be averaged over the 40 test items to obtain single measures of estimation accuracy. The second method was to arbitrarily establish a DIF criterion value for each true DIF index and then to observe the proportion of true positive and true negative DIF identifications or "hit rates" over replications and items. The DIF criteria used were $|\mu_{STD}| > .10$, and $|-\ln(\psi)| > \ln(2)$.

Regardless of the degree of incongruence, the simulated test overall was free from DIF, as measured by $\frac{1}{40}\Sigma(\mu_{STD})$ and $\frac{1}{40}\Sigma -\ln(\psi)$. These average population DIF values varied only from -.0116 to .0057 and from -.0409 to .0429, respectively, and indicated, on average, a test free from DIF.

Results

The results of the computer simulations are summarized in a series of plots given by Figures 1-3. Figures 1 and 2 show the results of the MH and STD estimators, respectively, in terms of average bias or [(-ln MH) - (-ln ψ)] (across items), average standard error (SE of estimate across items) and average root mean squared error (RMSE across items), each as a function of distributional overlap. In each of these figures, the solid lines represent those situations where the variances of *F* and *B* ability distributions were equal to 1.0; the dotted lines indicate those situations in which the variance of the ability distribution of *F* was .5 while the variance of the ability distribution of *B* remained at 1.0.

Insert Figures 1 and 2 About Here

Figure 3 shows the proportion of times (out of 100 replications) that MH and STD accurately identified items as either having no DIF or as having DIF, as measured by the DIF criteria given above. Because the simulated tests were, on average, free from DIF, these "hit rates" tended to be situations involving *true negatives* (i.e., items without DIF). Once again, the solid and dotted lines represented the two different variance conditions.

Insert Figure 3 Here

MH Results

It was anticipated that the bias in MH would become increasingly *negative* as overlap decreased, due to the effect of ordered distributions on MH, as discussed by Zwick (1990) (i.e., as $g_F(\theta)$ became less than $g_B(\theta)$ in terms of stochastic ordering, -ln MH should have favored *B* in terms of the DIF analysis). Figure 1 shows that this did not occur. The MH bias remained slightly positive but basically close to zero until the percentage of overlap fell to values near .1. The obvious explanation for the apparent unbiased behavior of MH even as overlap approached zero was due to the presence of empty cells or zero frequencies for many of the score categories. These zero contributions to the overall estimate of the log of the common odds-ratio did not affect the *no-DIF* conclusion. And because this was the true situation, the MH estimates appeared to be unbiased.

The instability of the MH estimator as the percentage of overlap decreased was also apparent from the increase in the SE. See Figure 1. Overall, the RMSE remained fairly constant until the percentage of distributional overlap was less than .4. This value of .4 represented mean differences of -1.75 in the equal variance case and -1.5 in the unequal variance case.

The correct identification of DIF and no-DIF items remained fairly high, above .90, for MH until overlap reached approximately .3. The MH Hit Rate fell off sharply after that. See Figure 3.

STD Results

Similar findings were noted for the STD estimator. Assuming that the stochastic ordering of the two distributions would once again produce results which (falsely) favored *B*, it was again anticipated that the bias in STD would become increasingly *negative* as overlap decreased. Figure 2 shows that the STD estimator remained fairly unbiased once again, even when the overlap percentage approached zero. The SE again increased as the two distributions separated, which resulted in an increase in the RMSE. These results were consistent across both variance conditions.

Correct DIF identification with STD was consistently lower than that of MH until the percentage of distributional overlap reached .2. After that, the situation was reversed with STD performing better than MH. See Figure 3.

Asymptotic Bias

In order to determine the effect of sample sizes on these results, it was possible to evaluate MH and STD as the number of items, k, remained fixed and the sample sizes within the cells of the $k+1$ 2 X 2 tables used to obtain MH and STD increased indefinitely. This was done analytically, using a recursive procedure to obtain $f(X|\theta)$ and hence, $h(X)$, for each group, as described in Lord and Wingersky (1984, p.454). This evaluation did two things. First, because the sample sizes were infinite, SE was driven to zero. And because the cells contained expected frequencies, zero cell frequencies were eliminated. These analytical values of MH and STD were then compared to μ_{STD} and - ln ψ. The difference between the analytical and the population value was termed *asymptotic bias*. Figure 4 shows the average asymptotic bias (over items) of MH and STD as functions of overlap. In this figure, the anticipated direction of the bias was confirmed. Both MH and STD were biased in the direction of *B* (i.e., negatively). Note that the severity of the bias for MH and STD was about the same. The appearance of differences between MH and STD in Figure 4 was due to differences in the scales of the two estimators.

Insert Figure 4 Here

Figures 1, 2, and 4 illustrate an interesting paradox in using the MH and STD estimators when the two ability distributions were non-overlapping. The statistics remained fairly unbiased *for tests with no DIF* when the sample sizes were moderate, due to the many zero cell frequency contributions to MH and STD. However, for these moderate sample sizes, the SE was fairly substantial. The end result was that the RMSE increased as overlap decreased. Increasing the sample sizes would certainly decrease the SE of the MH and STD estimates *but* coincidentally it would increase the bias. The net result would be the same, namely that the RMSE would increase as the percentage of overlap decreased.

Results for Completely Congruent Cases

Three other simulations were conducted to show the effects of reduced test reliability alone, as opposed to distributional incongruence, on DIF identification. These simulations were conducted so that the variances of both latent ability distributions were 1.0, but $\mu_{F\theta}$ and $\mu_{B\theta}$ were both set at -1.0, -2.0, and -3.0. The results were then compared to the original case of complete congruence, with $\mu_{F\theta}$ and $\mu_{B\theta}$ equal to 0.0, as well as the non-overlapping cases illustrated previously in Figures 1-4. In this way the effects due to distributional incongruence could be somewhat separated from those due only to reduced reliability. These results are summarized in Table 1.

Insert Table 1 Here

As Table 1 illustrates, most of the increases in SE (and, consequently in RMSE) and the decreases in Hit Rates seen in Figures 1-4 were due to distributional incongruence rather than to lowered test reliabilities alone. As long as the two distributions remained congruent, SE and Hit Rates were fairly consistent. And although there was some decline in DIF identification performance as reliability decreased, it was not as severe as that observed when overlap was less than 1.0. However, it should be pointed out that reduced test reliability and distributional incongruence, as modeled in these computer simulations, were somewhat confounded. Obviously, it was impossible to shift $g_F(\theta)$ too far in the negative direction without affecting the test reliability of *F*, due to the nonzero lower asymptote imposed by the three-parameter logistic function.

Although distributional incongruence imposed a reduced reliability condition on *F*, it was necessary to tolerate this confounding unless the item parameters were modified across each simulation condition, which was an unappealing alternative.

Table 1 also shows that the average asymptotic bias remained relatively unchanged as long as the two distributions were congruent. The average bias for samples of 500 was fairly close to the asymptotic results, even when test reliability was reduced.

Discussion

Although the results of these simulations were obtained using item response models estimated from a specific test and abilities generated from specific distributions, it is believed that these results are generalizable to a broader class of testing situations because of the wide range of distributional incongruence studied and because the test that was used to generate the responses was typical of many achievement tests. The major conclusion drawn from this study was that the use of the observed score, *X*, as a latent ability surrogate in computing MH and STD appeared to be acceptable, even when the degree of distributional incongruence was fairly substantial. DIF identification by MH and STD was acceptable for latent ability distributions that were as much as 1.5 to 2.0 standard deviations apart.

These results would appear to hold for tests which contain few DIF items. A similar study should be conducted to investigate the effect of the severity of distributional incongruence on tests where the occurrence of DIF is more frequent.

References

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*, 355-368.

Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine, 1*, 217-227.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure.* Program Statistics Research Technical Report No. 86-89. Princeton, NJ: Educational Testing Service.

Kendall, M., Stuart, A., & Ord, J.K. (1987). *Kendall's advanced theory of statistics* (Vol. 1) (5th ed.). New York: Oxford University Press.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 453-461.

Mislevy, R. J., & Bock, R. D. (1989). *PC-Bilog 3* [Computer program]. Moorseville, IN: Scientific Software, Inc.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

Footnote

[1] Item parameter estimates are not included in this paper but will be provided upon request.

Table 1

DIF Results for Completely Congruent Cases

Congruent Base & Focal Distributions (μ, σ^2)	Reliabilities (Base, Focal)	MH			STD		
		ASYMPTOTIC BIAS	BIAS SE RMSE	HIT RATES	ASYMPTOTIC BIAS	BIAS SE RMSE	HIT RATES
(0, 1)	(.91, .93)	.0039	.0125	.967	.0011	.0031	.957
			.1568			.0267	
			.1647			.0279	
(-1, 1)	(.86, .86)	.0320	.0432	.975	.0059	.0142	.969
			.1692			.0273	
			.1761			.0311	
(-2, 1)	(.69, .65)	.0152	.0138	.941	.0021	.0068	.928
			.1973			.0246	
			.1905			.0254	
(-3, 1)	(.34, .28)	-.0049	-.0332	.885	-.0006	-.0027	.921
			.2463			.0229	
			.2282			.0227	

Figure Captions

Figure 1. Bias, SE, and RMSE as a function of overlap for MH

Figure 2. Bias, SE, and RMSE as a function of overlap for STD

Figure 3. Hit rates as a function of overlap

Figure 4. Asymptotic bias as a function of overlap

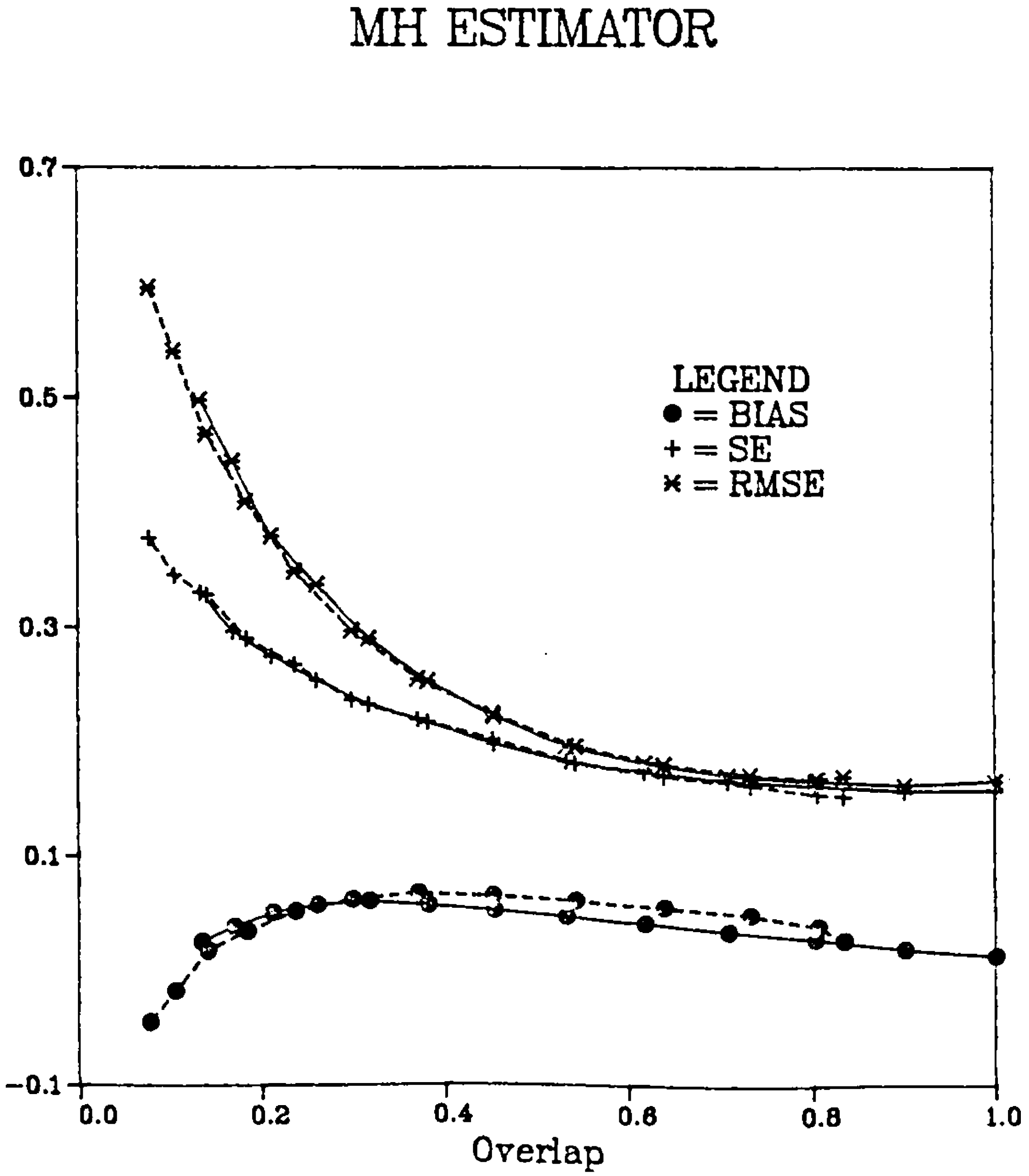

Figure 1. Bias, SE, and RMSE as a function of overlap for MH

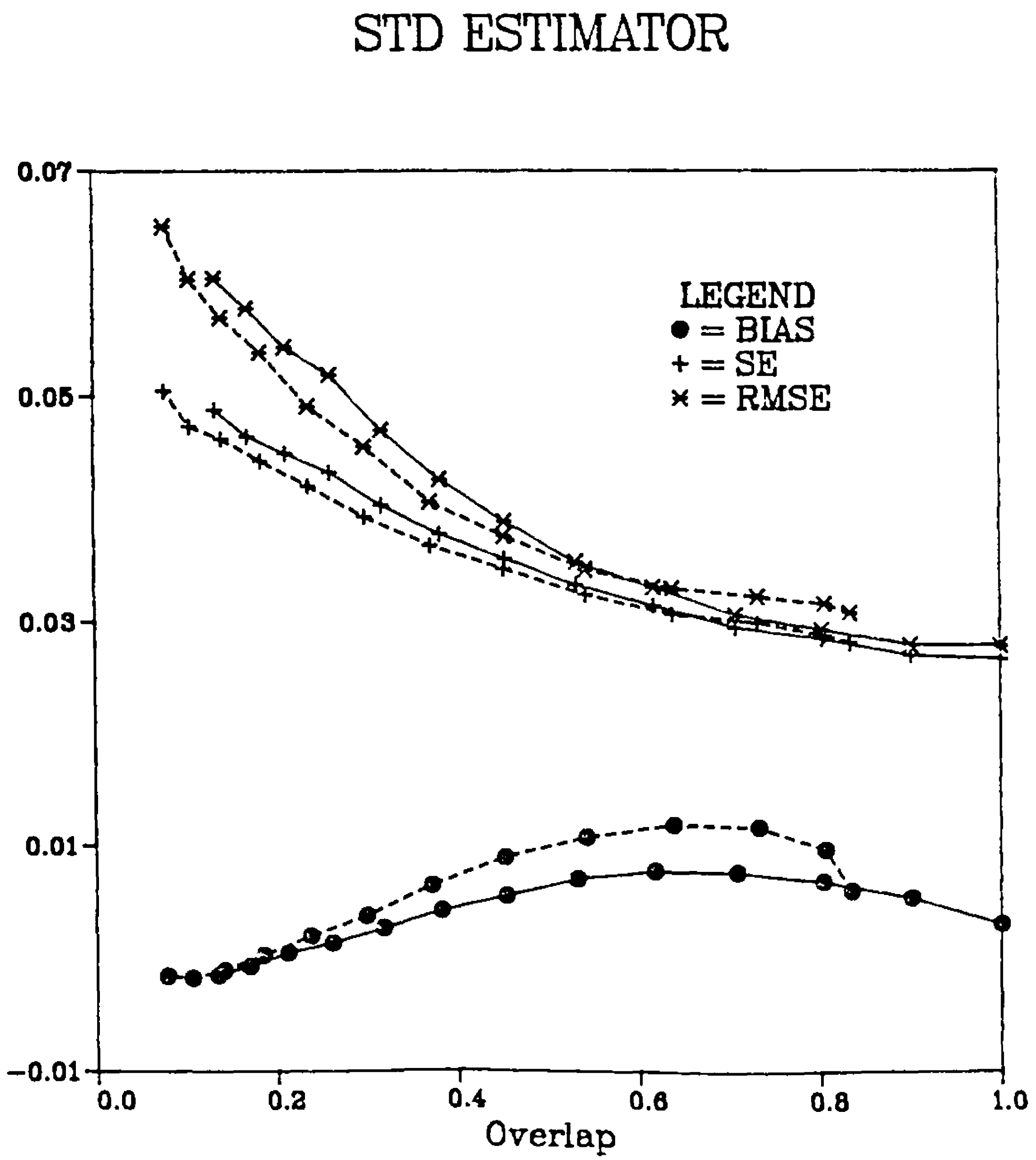

Figure 2. Bias, SE, and RMSE as a function of overlap for STD

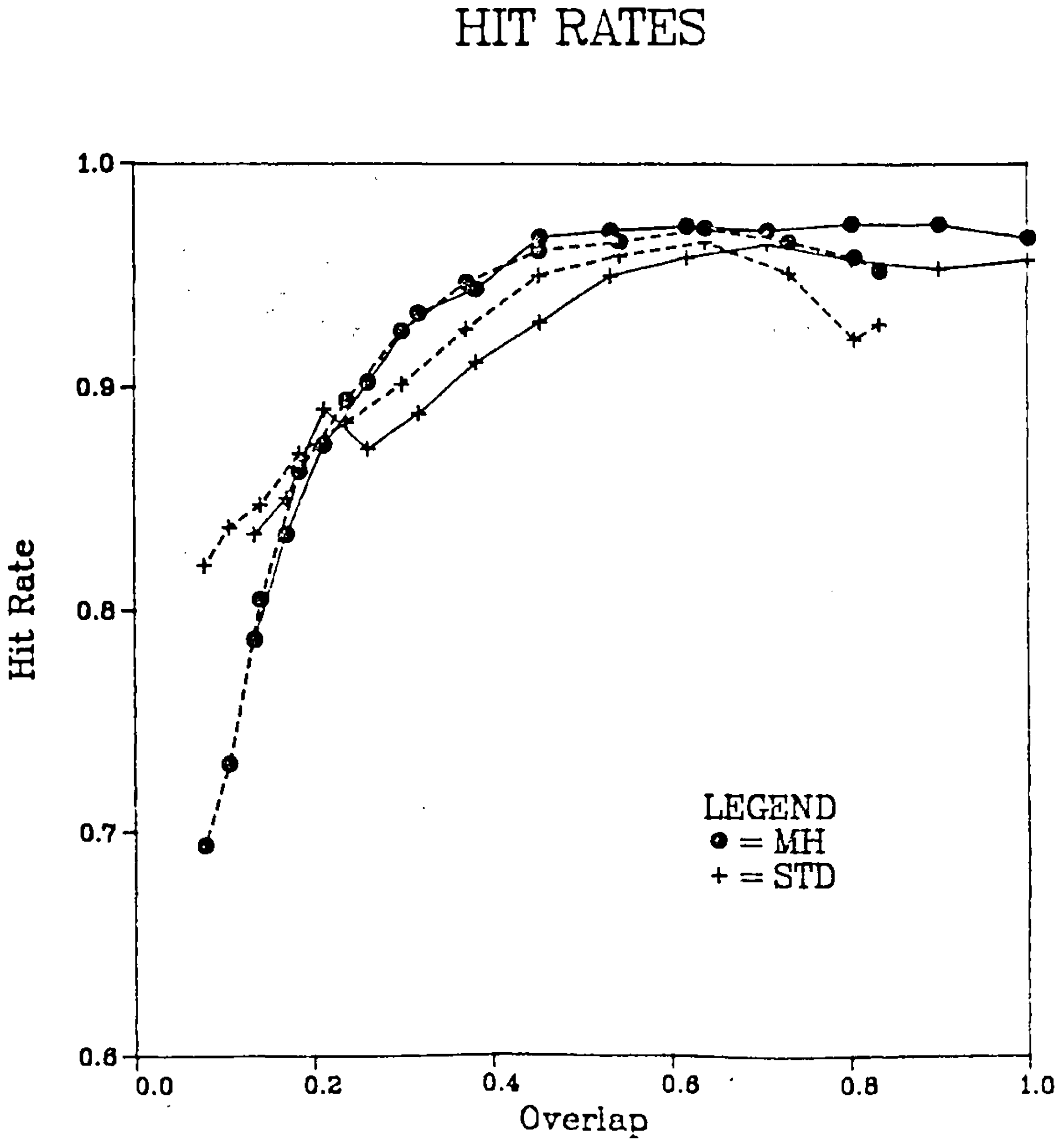

Figure 3. Hit rates as a function of overlap

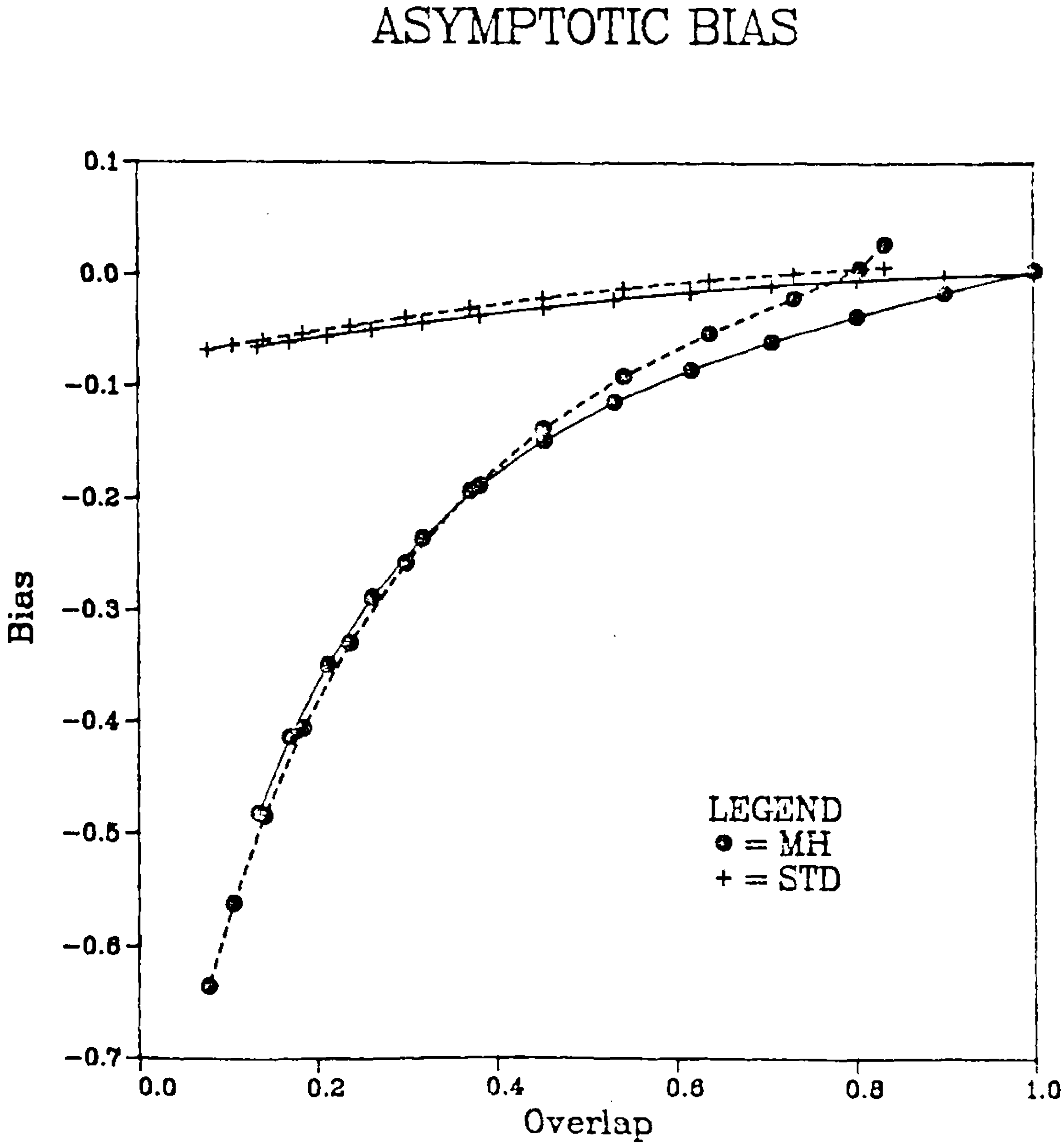

Figure 4. Asymptotic bias as a function of overlap