DOCUMENT RESUME

ED 382 658 TM 023 092

AUTHOR	Zwick, Rebecca
TITLE	The Effect of the Probability of Correct Response on the Variability of Measures of Differential Item Functioning. Program Statistics Research Technical Report No. 94-4.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-44
PUB DATE	Aug 94
NOTE	21p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Evaluation Methods; *Item Bias; Measurement Techniques; *Probability; Simulation; *Test Items
IDENTIFIERS	*Mantel Haenszel Procedure; Rasch Model; *Variability; Variance (Statistical)

ABSTRACT

The Mantel Haenszel (MH; 1959) approach of Holland and Thayer (1988) is a well-established method for assessing differential item functioning (DIF). The formula for the variance of the MH DIF statistic is based on work by Phillips and Holland (1987) and Robins, Breslow, and Greenland (1986). Recent simulation studies showed that the MH variances tended to be larger when items were administered to "examinees" who were randomly selected from a population than when items were administered adaptively. An analytic perspective shed some light on this puzzling result. Although the general form of the MH variance is complex and does not provide an intuitive understanding of the phenomenon, application of certain Rasch model assumptions yields a simple expression that appears to explain the difference in variances for adaptive versus nonadaptive administration. One table, two figures. (Contains 13 references.) (Author)

RR-94-44

The Effect of the Probability of Correct Response on the Variability of Measures of Differential Item Functioning

Rebecca Zwick
Educational Testing Service

PROGRAM STATISTICS RESEARCH

Technical Report No. 94-4

Educational Testing Service
Princeton, New Jersey 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

The Effect of the Probability of Correct Response on the Variability of Measures of Differential Item Functioning

Rebecca Zwick
Educational Testing Service

Program Statistics Research
Technical Report No. 94-4

Research Report No. 94-44

Educational Testing Service
Princeton, NewJersey 08541

August 1994

July 24, 1994

The Effect of the Probability of Correct Response on the Variability of Measures of Differential Item Functioning

Rebecca Zwick

Educational Testing Service

Acknowledgments: I am grateful to Charlie Lewis, who provided valuable consultation and suggested the use of a Rasch model analysis to elucidate some puzzling simulation results, Dorothy Thayer and Marilyn Wingersky, who were my collaborators on the simulation research that prompted this work, and Neil Dorans and Nancy Petersen, who commented on the manuscript.

Requests for reprints should be sent to Rebecca Zwick, Educational Testing Service, Rosedale Road, Princeton, NJ 08541.

Abstract

The Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) is a well-established method for assessing differential item functioning (DIF). The formula for the variance of the MH DIF statistic is based on work by Phillips and Holland (1987) and Robins, Breslow and Greenland (1986). Recent simulation studies showed that the MH variances tended to be larger when items were administered to "examinees" who were randomly selected from a population than when items were administered adaptively. An analytic perspective shed some light on this puzzling result. Although the general form of the MH variance is complex and does not provide an intuitive understanding of the phenomenon, application of certain Rasch model assumptions yields a simple expression that appears to explain the difference in variances for adaptive versus nonadaptive administration.

Keywords: Differential item functioning, Mantel-Haenszel, item bias, Rasch model

The Effect of the Probability of Correct Response on the Variability of Measures of Differential Item Functioning

Overview

The Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) is a well-established method for assessing differential item functioning (DIF). The MH index of DIF, *MH D-DIF*, is proportional to the natural log of the MH odds ratio estimate, $\hat{\alpha}_{\text{MH}}$. Phillips and Holland (1987) used a new approach to derive an estimated standard error for $\ln(\hat{\alpha}_{\text{MH}})$; their result proved to be identical to that of Robins, Breslow and Greenland (1986).

Recent simulation studies (Zwick, Thayer, & Wingersky, in press; 1994) showed a puzzling result: The standard error of *MH D-DIF* tended to be larger when items were administered to "examinees" who were randomly selected from a population than when items were administered adaptively; that is, using an algorithm that selects items with difficulty levels close to the examinee's ability.

The general form of the Phillips-Holland variance formula is quite complex and does not lend itself to an intuitive understanding of this phenomenon. However, if DIF is assumed to be absent and if the item response functions are assumed to follow the Rasch model, the variance takes on a simple form that appears to explain the difference in standard errors in adaptive versus nonadaptive administration.

Mantel-Haenszel DIF Analysis

In the MH approach, examinees are first grouped on the basis of a matching variable that is intended to be a measure of ability in the area of interest. In typical DIF applications, the matching variable is the total score on the test in which the item under investigation--*the studied item*--is embedded.

The score on the studied item, group membership, and the value of the matching variable for each examinee define a 2 x 2 x K cross-classification of examinee data, where K is the number of levels of the matching variable. Assume that there are T_k examinees at the *k*th levrl of the matching variable. Of these, n_{Fk} are in the group of interest--*the focal group*--and n_{Rk} are in the comparison group, or *reference group*. Of the n_{Rk} reference group members, assume that A_k answered the studied item correctly while B_k did not. Similarly C_k of the n_{Fk} matched focal group members answered the studied item correctly, whereas D_k did not.

Within the *k*th level of the matching variable, let the ratio of the odds of answering the item correctly for the reference group to the corresponding odds for the focal group be defined as α_k, $k = 1, 2, ..., K$. The MH χ^2 test approximates the uniformly most powerful unbiased test of the hypothesis,

$$H_0: \alpha_k = \alpha = 1, \quad k = 1, 2, ..., K \tag{1}$$

versus the alternative,

$$H_1: \alpha_k = \alpha, \ \alpha \neq 1, \quad k = 1, 2, ..., K \tag{2}$$

(Holland & Thayer, 1988).

The MH measure of DIF is

$$MH\,D - DIF = -2.35\ln(\hat{\alpha}_{\mathrm{MH}}) \tag{3}$$

where $\hat{\alpha}_{\mathrm{MH}}$ is the Mantel-Haenszel conditional odds-ratio estimator given by

$$\hat{\alpha}_{\mathrm{MH}} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}$$

In equation 3, the transformation of $\hat{\alpha}_{\mathrm{MH}}$ places *MH D-DIF* on the ETS delta scale of em difficulty (Holland & Thayer, 1985). The effect of the minus sign is to make *MH D-DIF* negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group.

The variance of $\ln(\hat{\alpha}_{\mathrm{MH}})$ is given by

$$\mathrm{Var}(\ln(\hat{\alpha}_{\mathrm{MH}})) = \frac{1}{2M^2}\sum_{k=1}^{K}\frac{1}{T_k^2}E\left[\left(A_kD_k + \alpha\, B_kC_k\right)\left(A_k + D_k + \alpha\left(B_k + C_k\right)\right)\right], \qquad (4)$$

where

$$M = E(\sum_{k=1}^{k} A_kD_k / T_k) \qquad (5)$$

(see Phillips & Holland, 1987, their equations 4 and 8).[1] The sample estimate is

$$\mathrm{V\hat{a}r}(\ln(\hat{\alpha}_{\mathrm{MH}})) = \frac{\sum_k U_k\, V_k / T_k^2}{2\,(\sum_k A_k\, D_k / T_k)^2}, \qquad (6)$$

where $U_k = (A_k\ D_k) + \hat{\alpha}_{\mathrm{MH}}(B_k\ C_k)$ and $V_k = (A_k + D_k) + \hat{\alpha}_{\mathrm{MH}}(B_k + C_k)$ (Phillips & Holland, 1987, their equation 9). The estimated standard error for *MH D-DIF* can then be expressed as

$$SE(MH\,D - DIF) = 2.35\sqrt{\mathrm{V\hat{a}r}(\ln(\hat{\alpha}_{\mathrm{MH}}))}$$

(Holland & Thayer, 1988).

Motivation for the Derivation of a Simplified Form for $\mathrm{Var}(\ln(\hat{\alpha}_{\mathrm{MH}}))$

In the simulation study of Zwick, Thayer, and Wingersky (in press), item responses were generated using a 3PL model. In the main portion of the study, which involved computer-adaptive tests (CATs), 25 items o:1t of a pool of 75 were "administered" to each examinee using an algorithm that selected the most informative item at the examinee's current estimate of ability. Ability was reestimated following each item response. A separate portion of the study investigated DIF results for items that were administered nonadaptively. A comparison showed that *SE(MH D-*

DIF) tended to be larger for nonadaptive than for CAT items, although the sets of items had similar generating parameters and DIF properties. For example, for n_R = *900* reference group members and n_F = *100* focal group members, *SE(MH D-DIF)* had a range of 0.6 to 1.1 for the nonadaptive items, compared to 0.5 to 0.7 for the CAT items.

To explore this issue further, both nonadaptive and CAT DIF results were obtained for the items in the CAT pool. Figures 1 and 2 (from Zwick, Thayer, & Wingersky, 1994) show these *MH D-DIF* statistics and their standard errors, respectively. (There were 71, rather than 75, items because four items were never administered in the simulated CAT.) The values plotted along the horizontal axis are based on nonadaptive administration to n_R = 900 reference group members and n_F = 100 focal group members. The values plotted along the vertical axis are based on only the examinees who received the item in a CAT administration. The method used to estimate the CAT-based *MH D-DIF* statistics and their standard errors for n_R = 900, n_F = 100 is described in Zwick, Thayer, and Wingersky (in press). The same matching variable, an item response theory (IRT)-based expected true score, was used for the CAT and nonadaptive DIF statistics.

Insert Figures 1-2 about here.

Figure 1 shows that the *MH D-DIF* statistics are clustered around the 45-degree line; there were no systematic differences for the two types of administration. The standard errors, however were substantially different, with the nonadaptive standard errors exceeding the CAT standard errors by an average of 15 percent. Figure 2 shows that the nonadaptive standard errors were larger for almost every item.

To determine whether these findings were related to the particular CAT algorithm used in this study, another analysis was conducted, comparing nonadaptive administration to a "pseudo-CAT" administration, in which examinees were eliminated from the DIF analysis of a particular item if their abilities departed subs antially from the estimated item difficulty. Results were very similar to those obtained in Figures 1 and 2, suggesting that this phenomenon was associated with the ability range of the examinees, but was not unique to the implemented CAT algorithm. A later study by Way (1994), which used a different CAT algorithm, revealed the same phenomenon.

MH D-DIF and its Standard Error under the Rasch Model

In the Rasch model, the probability of correct response in group G can be expressed as

$$P_G(X=1\,|\,\theta)=\left\{1+\exp\left[-(\theta-b_G)\right]\right\}^{-1}, \tag{7}$$

where X is the score on the item, X = (0, 1), with "1" indicating a correct response, θ represents ability, and b_G is the item difficulty in group G (G=R, F). Although a model that ignores guessing and treats items as equally discriminating cannot be expected to hold for typical multiple-choice tests, the Rasch model often proves useful for explanatory purposes. For example, Holland and Thayer (1988) offered a Rasch-based analysis that elucidates the relation between *MH D-DIF* and the item difficulty parameters for the reference and focal groups. From an IRT perspective, DIF can be defined as a difference in item response functions(IRFs) for two groups (Lord, 1980). Holland and Thayer (1988) showed that under certain conditions, identity of item response functions across groups for the studied item satisfies the MH null hypothesis (equation 1) and a difference in IRFs across groups corresponds to the MH alternative hypothesis (equation 2). The required conditions are:

(i) within each of the groups (reference and focal), the IRFs follow the Rasch model in (7),

(ii) the matching variable is the number-right score based on all items, *including the studied item*, and

(iii) the items have the same IRFs for the two groups, with the possible exception of the studied item.

Under these conditions, the odds ratios α_k in (1) and (2) are equal to $\exp(b_F - b_R)$, where b_F and b_R are the item difficulties for the reference and focal groups, respectively. The quantity $\exp(b_F - b_R)$ is constant across all levels of the matching variable and is equal to one when the reference and focal groups have the same IRF. Zwick (1990) showed that the correspondence between IRF and MH definitions of DIF could not be assured to hold for a more general class of item response models that includes the usual two- and three-parameter logistic (2PL and 3PL) models.[2] Nevertheless, in simulation studies, this result has been found to hold approximately under moderate departures from (7). Because the Rasch model has provided useful insights about the behavior of the *MH D-DIF* statistic, even when the true model is more complex, it seemed worthwhile to consider a Rasch-based analysis of *SE(MH D-DIF)* in attempting to explain the surprising finding about MH standard errors.

The result of applying Rasch model and no-DIF assumptions to the Phillips and Holland (1987) variance in equation 4 (above) will now be demonstrated. As a first step, the expected table frequencies will be obtained under the following simplifying assumptions:

(1) The matching variable is S, the total test score, including the studied item.

(2) IRFs for the two groups can be represented by a single Rasch model (as in 7). This further implies that

(2a) Conditional independence holds, i.e., $P(X=x\mid\theta)=\prod_i P(X_i=x_i\mid\theta)$ for all θ, where i indexes items.

(2b) There is no DIF; that is, $P_R(X=1\mid\theta)=P_F(X=1\mid\theta)$ holds for all items.

Assumption 2b may seem unduly restrictive, but, in fact, simulation results show that the size of the MH standard errors depends very little on the true magnitude of DIF (e.g., Zwick, Thayer, & Wingersky, in press; 1994).

To obtain the expected table frequencies under these assumptions, a useful property of the Rasch model can be exploited: Under this model, S is sufficient for θ and therefore, $P_G(X=1\mid S,\theta)=P_G(X=1\mid S)$ (Zwick, 1990). Invoking this property, as well as the above assumptions, it can be shown that $P_R(X=1\mid S)=P_F(X=1\mid S)$. Note that identity of item response functions for the reference and focal groups (assumption 2b) does not, in general, imply that $P_R(X=1\mid S)=P_F(X=1\mid S)$. For example, this implication would not hold for the usual 2PL and 3PL models.

The expected frequencies in stratum k of the 2 x 2 x K table are as shown in Table 1, where $\pi_k = P_{\mathrm{R}}(X=1 \mid S=s_k)=P_{\mathrm{F}}(X=1 \mid S=s_k)$ is the probability of a correct response in stratum k, and $n_{\mathrm{R}k}$ and $n_{\mathrm{F}k}$ are the reference and focal group frequencies, respectively, in stratum k.

Insert Table 1 about here

In deriving their variance formula, Phillips and Holland (1987, equation 4 above) assumed that A_k and C_k follow the independent binomial distributions, $A_k \sim B(n_{\mathrm{R}k},\pi_k)$ and $C_k \sim B(n_{\mathrm{F}k},\pi_k)$. (Alternatively, it can be assumed that A_k is a hypergeometric variate.) In the present context, the assumption of binomial distributions is redundant, since the Rasch model assumptions imply that, conditional on T_k, $A_k \sim B(n_{\mathrm{R}k},\pi_k)$ and $C_k \sim B(n_{\mathrm{F}k},\pi_k)$ (Rasch, 1960, p. 180).

Now, because Assumption 2b implies that $\alpha=1$, equation 4 can be simplified as follows:

$$\mathrm{Var}(\ln(\hat{\alpha}_{\mathrm{MH}}))=\frac{1}{2M^2}\sum_{k=1}^{k}\frac{1}{T_k}E(A_kD_k+B_kC_k).\ ,$$

where M is given by equation 5. Invoking the independence of the two binomials yields

$$\mathrm{Var}(\ln(\hat{\alpha}_{\mathrm{MH}}))=\frac{1}{2M^2}\sum_{k=1}^{K}\frac{1}{T_k}\left[E(A_k)E(D_k)+E(B_k)E(C_k)\right],$$

where M can now be expressed as $M=\sum_{k=1}^{k}\frac{1}{T_k}E(A_k)E(D_k)$.

The expected cell frequencies from Table 1 can now be substituted, leading to the result

$$\begin{aligned}\mathrm{Var}(\ln(\hat{\alpha}_{\mathrm{MH}}))&=M^{-1}\\&=\left\{\sum_{k=1}^{K}\left(\frac{n_{\mathrm{R}k}n_{\mathrm{F}k}}{n_{\mathrm{R}k}+n_{\mathrm{F}k}}\right)\pi_k(1-\pi_k)\right\}^{-1}.\end{aligned} \qquad (8)$$

The expression in (8) is a theoretical formulation of the variance of $\ln(\hat{\alpha}_{\mathrm{MH}})$ under the stated Rasch assumptions.[3]

Note the obvious similarity of the result in (8) to the asymptotic variance of the logit of a sample proportion, $\hat{p}$, which is given by

$$\mathrm{Var}(\mathrm{logit}(\hat{p})) = \mathrm{Var}(\ln(\hat{p}/(1-\hat{p})) = (n\pi(1-\pi))^{-1}, \qquad (16)$$

where π is the population proportion (Agresti, 1990).[4] When π = .5, the variance of $\mathrm{logit}(\hat{p})$ is minimized. Similarly, for fixed values of K, $n_{\mathrm{R}k}$ and $n_{\mathrm{F}k}$, the variance in (8) is minimized if the probability of a correct response in each stratum is equal to .5. The condition in which $\pi_k = .5$ for all k is consistent with CAT administration, which provides for items to be assigned to examinees of an appropriate, and thus rather narrow, ability range. Large departures from this condition occur for nonadaptive items. These findings appear to shed light on the differences in the MH standard errors fnr adaptive and nonadaptive administration.

Two subtle aspects of these results are worthy of note. First, comparison of the MH variances for adaptive and nonadaptive items is somewhat more complicated than the exposition above suggests. The variances for these two types of items differ not only in terms of the values of π_k , but in terms of the values of $n_{\mathrm{R}k}$ and $n_{\mathrm{F}k}$ and possibly K, the number of strata. In an adaptive administration, examinees are typically concentrated in a smaller number of test score levels. Second, whereas the estimated MH variances are related to the item proportion correct (the classical item difficulty), they do not have a straightforward relationship to the IRT difficulty parameter, b. For a given value of b, the proportion correct for an adaptive item will tend to be closer to .5 than the proportion correct for a nonadaptive item. Therefore, MH standard errors will tend to be larger for a nonadaptive item than for an adaptive item with the same b value.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Fischer, G. H. (1993). Notes on the Mantel-Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika*, *7*, 88-100.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty.* ETS Research Report No. 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity,* pp. 129-145. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, *43*, 425-431.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen and Lydiche.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, *42*, 311-323.

Way, W. D. (1994). *A simulation study of the Mantel-Haenszel procedure for detecting DIF with the NCLEX using CAT.* Technical report, Educational Testing Service.

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

Zwick, R., Thayer, D. T., & Wingersky, M. (in press). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement.*

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). *DIF analysis for pretest items in computer-adaptive testing.* ETS Research Report No. 94-33. Princeton, NJ: Educational Testing Service.

Footnotes

[1] The variance of $\ln(\hat{\alpha}_{\text{MH}})$ is infinite since $\ln(\hat{\alpha}_{\text{MH}})$ can take on the values ±∞ with positive probability. Strictly speaking, the Phillips-Holland variance formula applies not to $\ln(\hat{\alpha}_{\text{MH}})$, but to a Taylor series approximation to $\ln(\hat{\alpha}_{\text{MH}})$ that does have a finite variance. This qualification applies to all references to $\text{Var}(\ln(\hat{\alpha}_{\text{MH}}))$ in this paper.

[2] Fischer (1993) showed that the definitions of DIF agree for very restrictive cases of the 2PL. Among other requirements, it is necessary that the discrimination parameters be known and that they be the same for the reference and focal groups on each item.

[3] Although this need not have been the case, substitution of the expected cell counts of Table 1 for their sample counterparts in the *estimated* variance formula of equation 6 produces the same result as (8). Without the more rigorous derivation, however, it would be difficult to know how to interpret the obtained expression.

[4] This variance formula is again based on Taylor series appıoximation; see footnote 1.

Table 1

Expected Frequencies Under Rasch Model Assumptions

	Response		
Group	1	0	Total
Reference	$E(A_k)=n_{Rk}\pi_k$	$E(B_k)=n_{Rk}(1-\pi_k)$	n_{Rk}
Focal	$E(C_k)=n_{Fk}\pi_k$	$E(D_k)=n_{Fk}(1-\pi_k)$	n_{Fk}
Total	$T_k\pi_k$	$T_k(1-\pi_k)$	T_k

Figure 1

MH D-DIF for 71 Items ($n_R = 900$, $n_F = 100$)

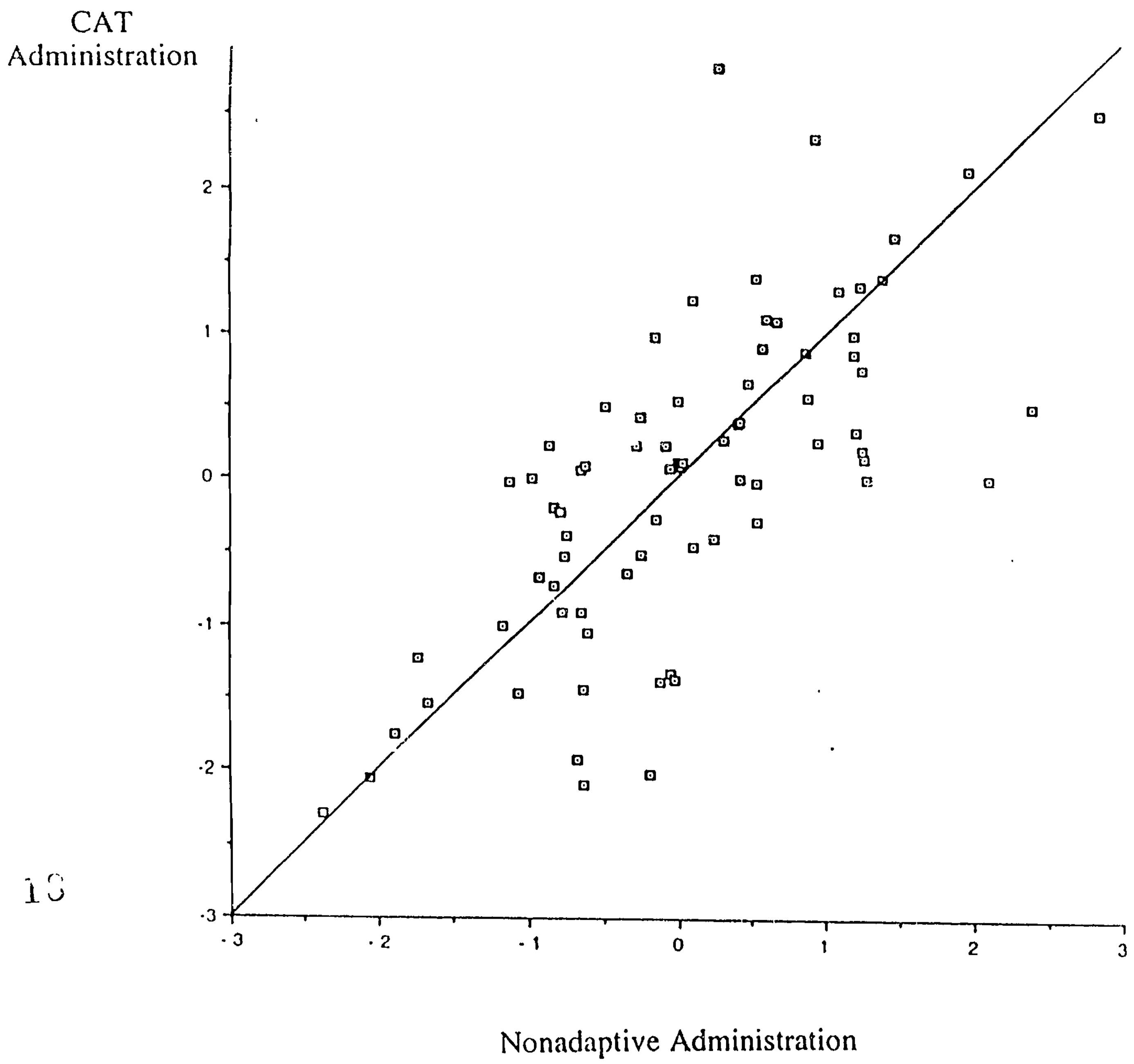
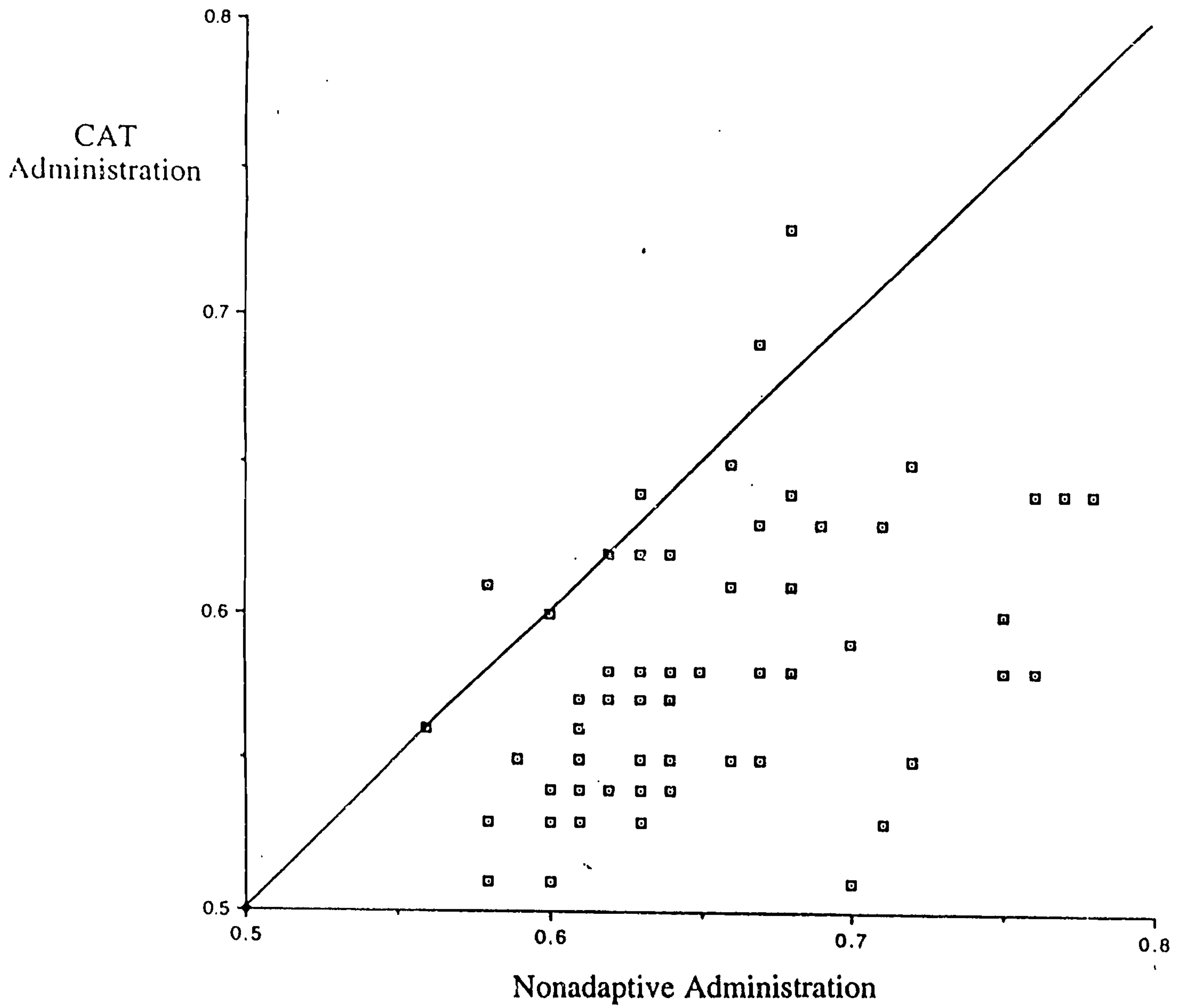


Figure 2

SE(MH D-DIF) for 71 Items ($n_R = 900$, $n_F = 100$)


CAT
Administration
Nonadaptive Administration