A Simulation Study of Methods for Assessing Differential Item Functioning in Computerized Adaptive Tests

Rebecca Zwick, Dorothy T. Thayer, and Marilyn Wingersky

Educational Testing Service

Simulated data were used to investigate the performance of modified versions of the Mantel-Haenszel method of differential item functioning (DIF) analysis in computerized adaptive tests (CATs). Each simulated examinee received 25 items from a 75-item pool. A three-parameter logistic item response theory (IRT) model was assumed, and examinees were matched on expected true scores based on their CAT responses and estimated item parameters. The CAT-based DIF statistics were found to be highly correlated with DIF statistics based on nonadaptive administration of all 75 pool items and with the true magnitudes of DIF in the simulation. Average DIF statistics and average standard errors also were examined for items with various characteristics. Finally, a study was conducted of the accuracy with which the modified Mantel-Haenszel procedure could identify CAT items with substantial DIF using a classification system now implemented by some testing programs. These additional analyses provided further evidence that the CAT-based DIF procedures performed well. More generally, the results supported the use of IRT-based matching variables in DIF analysis. *Index terms: adaptive testing, computerized adaptive testing, differential item functioning, item bias, item response theory.*

Many large-scale testing programs are now piloting or implementing computerized adaptive tests (CATs). These include the Scholastic Assessment Tests (SAT), the Graduate Record Examinations (GRE), the College Board Computerized Placement Tests, Praxis (successor to the NTE teacher assessment), and NCLEX (the licensure exam of the National Council of State Boards of Nursing), all developed by Educational Testing Service (ETS) and its client organizations; the COMPASS placement tests produced by the American College Testing Program; the Differential Aptitude Tests published by the Psychological Corporation; and the Armed Services Vocational Aptitude Battery (ASVAB).

The item responses collected from an examinee in a CAT may be a small fraction of the data that would have been collected in a corresponding nonadaptive test. Furthermore, the items received by each examinee are a nonrandom subset of the available pool of items. The introduction of CATs requires, therefore, that new approaches be developed for analyzing item properties, including differential item functioning (DIF).

The present study investigated whether simple noniterative DIF analysis methods can accommodate the data collected in a CAT. There are two reasons that DIF detection may be more important for CATs than it is for nonadaptive tests. First, because fewer items are administered in a CAT, each item response plays a more important role in the examinee's test score than it would in a nonadaptive testing format. Any flaw in an item, therefore, may be more consequential for the examinee. Second, administration of a test by computer creates several potential sources of DIF that are not present in conventional tests, such as differential computer familiarity, facility, and anxiety, and differential preferences for computerized administration. (See Powers & O'Neill, 1992, for a review of literature on this topic.) Legg & Buhr (1992) and Schaeffer, Reese, Steffen, McKinley, & Mills (1993) reported ethnic and gender group differences in some of these attributes. Their

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 2, June 1994, pp. 121-140

0146-6216/94/020121-20$2.25

findings suggest that attitudes toward computer testing may be surprisingly complex. For example, Schaeffer et al. found that Asian test-takers were most likely to have a computer available at home and were most likely to report that using the computer mouse was very easy. Yet both Schaeffer et al. and Legg & Buhr found that Asian examinees were more likely than any other ethnic group to state that they preferred paper-and-pencil to computerized administration.

Method

This study was an evaluation of the feasibility of conducting DIF analyses using modified versions of the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) approach of Holland & Thayer (1988) and the standardization method of Dorans & Kulick (1986). (Results of the standardization approach, which closely paralleled the MH findings, are discussed in Zwick, Thayer, & Wingersky, 1993.) Responses to three different pools of 75 items were simulated. In Pool 1, the items had no DIF; in Pool 2, the items had DIF that was uncorrelated with item difficulty; and in Pool 3, the items had DIF that was correlated with item difficulty. DIF analyses are typically based on two groups—the group of primary interest, or focal group, and the group to which the focal group is compared (the reference group). The only kind of DIF that was investigated here was a difference in item difficulty for the reference and focal groups, often called uniform DIF. The distance between reference and focal group means and the sample sizes for the two groups were varied, as were the DIF status of the items and the item difficulties and discriminations.

Using a CAT algorithm based on item information, each simulee (simulated examinee) was assigned 25 items from one of the three pools of 75 items. For DIF analysis, simulees were matched on the expected true score for the entire 75-item pool, computed using estimated trait (θ) values from the 25 CAT items and estimated item parameters [an approach suggested by Steinberg, Thissen, & Wainer (1990)].

To disentangle the effects of assigning items using the CAT algorithm and matching simulees on expected true score, a "nonadaptive control" analysis was conducted in which the matching variable for DIF analysis was the expected true score computed using a θ estimate based on responses to all 75 pool items. The results of this analysis were compared to the results obtained by matching on the CAT-based expected true score and to results obtained by matching on number-correct score, as in conventional MH analysis.

Simulation Procedures

The guiding principle in developing the simulation design was to aim for some reasonable compromise between an approach that was realistic (in that it reflected the properties of an actual CAT) and one that was simple enough to yield useful, interpretable results. The design of the simulation had three main components: determination of the "administration" conditions, definition of the properties of the simulated CAT, and specification of the parameters of the CAT pool items.

Administration Conditions

18 datasets were created, each corresponding to a CAT administration. The administrations were defined by three factors: the properties of the item pool (3 levels), the θ distributions of the focal group (3 levels), and the sample sizes (2 levels). The properties of the resulting 18 datasets are summarized in Table 1.

Item pools. Pool 1 had no DIF. The purpose of Pool 1 was to allow investigation of the functioning of the DIF methods in the null case. Any conclusion of DIF for this pool would constitute a Type I error. Two types of DIF pools were included. Pool 2 had DIF that was uncorrelated with reference group item difficulty b_R, and Pool 3 had DIF that was positively correlated with b_R. The reason for investigating the effect of a correlation between DIF and difficulty is that estimates of item difficulty and DIF have been found to be positively related to an appreciable degree for some pairs of examinee groups (e.g., Kulick & Hu, 1989). The b_R, item discrimination (a), and pseudo-guessing (c) parameters were the same across all three pools of

Table 1
F Group Distribution, N, and Item Pool for the 18 Administration Conditions

Condition	F Group Distribution	N_F	N_R	Pool
1	N(−1,1)	100	900	1
2	N(−1,1)	500	500	1
3	N(−1,1)	100	900	2
4	N(−1,1)	500	500	2
5	N(−1,1)	100	900	3
6	N(−1,1)	500	500	3
7	N(0,1)	100	900	1
8	N(0,1)	500	500	1
9	N(0,1)	100	900	2
10	N(0,1)	500	500	2
11	N(0,1)	100	900	3
12	N(0,1)	500	500	3
13	N(.5,1)	100	900	1
14	N(.5,1)	500	500	1
15	N(.5,1)	100	900	2
16	N(.5,1)	500	500	2
17	N(.5,1)	100	900	3
18	N(.5,1)	500	500	3

items; only the DIF properties varied.

θ *distributions*. There were three focal group θ distributions: N(−1,1), N(0,1), and N(.5,1). In each case, the reference group had a N(0,1) distribution. The differences between reference and focal group means were selected to be representative of group differences encountered in ETS DIF analyses.

Group sample size conditions. Two sample size conditions were included: $N_R = 500, N_F = 500$, and $N_R = 900, N_F = 100$, where N_R and N_F are the sample sizes for the reference (R) and focal (F) groups, respectively. These sample size conditions were selected to be similar to those that occur in ETS analyses.

CAT Simulation

The 25 CAT item responses were generated with the three-parameter logistic model (3PLM) item response function (IRF; Birnbaum, 1968), using the true item parameters and θ values. (To allow additional analysis, simulee responses also were generated for all of the pool items not administered in the CAT.) As described below, DIF was simulated by causing the bs for the reference and focal groups to differ on certain items.

The CAT simulation was designed as a simplified version of actual CATs being developed at ETS. The CAT algorithm, which is based on the approach of Lord (1977), selected as the next item to be administered the most informative item at the maximum likelihood estimate (MLE) of θ computed from the items already administered (see Lord, 1980, p. 72, for the definition of item information). Estimates of item information and simulee θ values were computed using estimated item parameters.

The study used a fixed-length CAT of 25 items, similar in length to a single section of the SAT and GRE CATs. The size of each pool was set at 75 items, which is smaller than a typical CAT pool, to ensure that most items would be administered. Even with this small pool, four items were never administered because, at every θ level, there were at least 25 items that were more informative.

In a process similar to that used in actual CATs, the first item administered was selected randomly from the four most informative items at $\theta = 0$. The second item was selected randomly from the three most informative items at a θ of either -2 or $+2$, depending on whether the first item was answered incorrectly or correctly, respectively. A simulee continued to receive the most informative item at a θ of -2 or $+2$ until at least one correct and one incorrect answer was produced. At that point, the MLE of θ was computed. Each subsequent item was selected to be the most informative item at the simulee's MLE of θ (provided that it had not already been given to that simulee). θ was reestimated by maximum likelihood following each item response. Simulees for whom finite θ estimates could not be computed were assigned θ values of -10 and $+10$ for all-incorrect and all-correct, respectively, and -5 and $+5$ for other patterns that did not produce finite estimates.

Item Parameters

Within each of the 18 conditions, a and b_R were varied, as well as the item DIF parameters (d), representing the degree to which the reference and focal group bs differed. Multivariate normal distributions (one for each pool) were used to model the joint distribution of the DIF and item parameters, with a natural log transformation applied to the a parameter. Item parameter estimates from actual admissions test datasets were used to determine the marginal means and standard deviations (SDs) of the item parameters. The mean and SD were set to (–.15, .30) for $\ln(a)$ and (0,.15) for b_R. To simplify the simulation, c_j was set equal to .15 for all items.

Marginal mean and SD of the d distribution. The DIF parameter for item j was defined as $d_j = b_{jR} - b_{jF}$. Therefore, a value of d greater than 0.0 implied that an item was easier for the focal group than for the reference group; d less than 0.0 implied that the item was more difficult for the focal group. In Pool 1 (no DIF), the mean and SD of d were 0.0. The determination of the distribution of d in Pools 2 and 3 was based on both theoretical and empirical findings on the relation of the MH-based DIF statistic, $\text{MH}_{\text{D-DIF}}$, to d (Holland & Thayer, 1988).

Donoghue, Holland, & Thayer (1993) used the work of Holland & Thayer (1988) to show that, under the following Rasch model assumptions, $\text{MH}_{\text{D-DIF}}$ (see Equations 1 and 2 below) provides an estimate of $4ad$: (1) within each of the groups (R and F), the IRFs follow the Rasch model (obtained by setting $c_j = 0$ for all items and $a_j \equiv a$ for all items); (2) the matching variable is the number-correct score based on all items, including the item under analysis, referred to as the *studied item*; and (3) the items have the same IRFs for the reference and focal groups (i.e., $b_{jR} = b_{jF} \equiv b_j$), with the possible exception of the studied item. When Assumptions 1–3 do not hold, the population odds ratios will not, in general, be constant across number-correct score levels (see Zwick, 1990).

Therefore, for the 3PLM, it is not possible to derive a general expression for the quantity estimated by $\text{MH}_{\text{D-DIF}}$. To provide a basis for selecting an appropriate marginal mean and SD of d for Pools 2 and 3, the regression of $\text{MH}_{\text{D-DIF}}$ on ad was examined for several sets of simulated data. The multiplicative constants were found to be close to 3.0 and the additive constants were approximately 0.0. Using this result, a mean of 0 and SD of .3 for d were selected to produce realistic distributions of $\text{MH}_{\text{D-DIF}}$. Because of the theoretical finding that $\text{MH}_{\text{D-DIF}}$ is proportional to ad in the Rasch case, and the empirical finding that $\text{MH}_{\text{D-DIF}}$ was approximately proportional to ad in this 3PLM simulation, true DIF was defined as ad in interpreting the results of this study.

Intercorrelations among item and DIF parameters. Intercorrelations were determined partly by the specified properties of Pools 1, 2, and 3 and partly by the observed correlations in the admissions test data (using $\text{MH}_{\text{D-DIF}}$ as a proxy for d). Pool 1 had no DIF, so d was uncorrelated with $\ln(a)$ and with b_R. By design, d also was uncorrelated with b_R in Pool 2, which corresponds to findings for male-female DIF in Kulick & Hu (1989) and in the admissions test data. The correlation of b_R and d for Pool 3 was set equal to .40, which is representative of the observed correlations for white-black and white-Asian DIF analyses. In all three pools, the correlation of $\ln(a)$ and d was set to 0.0 and the correlation of $\ln(a)$ and b_R was set to .40.

Discretized multivariate normal approach. To facilitate the summarization and interpretation of results, intervals were defined around the following desired values of $\ln(a)$, b_R, and d:

1. $\ln(a)$: $-.3$ and 0.0 (corresponding to a values of .74 and 1);
2. b_R: $-1.95, -1.3, -.65, 0.0, .65, 1.3$, and 1.95; and
3. d: $-.70, -.35, 0, .35$, and .70 in Pools 2 and 3; $d = 0$ for all items in Pool 1.

The probabilities from the multivariate normal distribution for the pool in question were used to assign probabilities to the cells of the $2 \times 7 \times 5$ contingency table corresponding to these $\ln(a)$, b_R, and d combinations. Marginal probabilities for a and b_R were constrained to be the same for all three pools. The cell probabilities were multiplied by the desired number of items for the pool and then rounded to integer values. The a, b_R, and d parameters for the three pools of items are given in Table 2.

Item parameter estimation for the CAT. The item parameter estimates used for computing item information and θ estimates were obtained through an analogue to a paper-and-pencil test administration. A sample of 2,000 simulees was administered all 75 items, and the LOGIST program (Wingersky, 1983; Wingersky, Patrick, & Lord, 1988) was used to estimate the a, b, and c parameters. Because 2,000 is a typical sample size for such calibrations, this approach allowed for the incorporation of a realistic amount of estimation error. The estimated a, b, and c parameters are given in Zwick, Thayer, & Wingersky (1993).

To allow comparisons across simulation conditions, a single set of item parameter estimates was used. Because it was not possible to define a calibration sample that included members of all three focal groups in a manner that was realistic or useful, only members of the reference population were included in the calibration. Therefore, the obtained estimates of item information functions and simulee θs were based on an incorrect (no DIF) model for the focal group. This closely approximates the situation that arises in actual testing situations when the true IRFs are different for the two groups, but the focal group constitutes only a small proportion of the calibration sample. In this case, item parameter estimates are, for all practical purposes, estimates of the reference group parameters.

Mantel-Haenszel DIF Analysis

In the MH method (Holland & Thayer, 1988), examinees are first grouped on the basis of a matching variable that is intended to be a measure of the ability of interest. In many DIF applications, the matching variable is the number-correct score on the test in which the studied item is embedded. The score on the studied item, group membership (R or F), and the value of the matching variable for each examinee define a $2 \times 2 \times K$ 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 kth level of the matching variable. Of these, N_{R_k} are in the R group and N_{F_k} are in the F group. Of the N_{R_k} R group members, A_k answered the studied item correctly and B_k did not. Similarly, C_k of the N_{F_k} matched F group members answered the studied item correctly, whereas D_k did not. The MH measure of DIF then can be defined as

$$\text{MH}_{\text{D-DIF}} = -2.35\ln(\hat{\alpha}_{\text{MH}}), \tag{1}$$

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

$$\hat{\alpha}_{\text{MH}} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}. \tag{2}$$

In Equation 1, the transformation of $\hat{\alpha}_{\text{MH}}$ places $\text{MH}_{\text{D-DIF}}$ on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The minus sign makes $\text{MH}_{\text{D-DIF}}$ negative when the item is more difficult for members of the

Table 2
True Values of a, b_R, and d Parameters for Item Pools 1, 2, and 3
(d = 0.0 in Pool 1)

	Pools 1, 2, 3		Pool 2		Pool 3	
Item	a	b_R	d	ad	d	ad
1	.74	−1.95	−.35	−.26	−.70	−.52
2	.74	−1.95	−.35	−.26	−.35	−.26
3	.74	−1.95	0.00	0.00	−.35	−.26
4	.74	−1.95	0.00	0.00	0.00	0.00
5	.74	−1.95	.35	.26	.35	.26
6	.74	−1.30	−.35	−.26	−.70	−.52
7	.74	−1.30	−.35	−.26	−.35	−.26
8	.74	−1.30	0.00	0.00	−.35	−.26
9	.74	−1.30	0.00	0.00	0.00	0.00
10	.74	−1.30	.35	.26	0.00	0.00
11	.74	−1.30	.35	.26	.35	.26
12	.74	−.65	−.35	−.26	−.35	−.26
13	.74	−.65	−.35	−.26	−.35	−.26
14	.74	−.65	0.00	0.00	0.00	0.00
15	.74	−.65	0.00	0.00	0.00	0.00
16	.74	−.65	0.00	0.00	0.00	0.00
17	.74	−.65	.35	.26	.35	.26
18	.74	−.65	.35	.26	.35	.26
19	.74	0.00	−.70	−.52	−.35	−.26
20	.74	0.00	−.35	−.26	−.35	−.26
21	.74	0.00	0.00	0.00	0.00	0.00
22	.74	0.00	0.00	0.00	0.00	0.00
23	.74	0.00	.35	.26	0.00	0.00
24	.74	0.00	.35	.26	.35	.26
25	.74	0.00	.70	.52	.35	.26
26	.74	.65	−.35	−.26	−.35	−.26
27	.74	.65	0.00	0.00	0.00	0.00
28	.74	.65	0.00	0.00	0.00	0.00
29	.74	.65	.35	.26	.35	.26
30	.74	.65	.35	.26	.35	.26
31	.74	.65	.70	.52	.70	.52
32	.74	1.30	−.70	−.52	−.35	−.26
33	.74	1.30	−.35	−.26	0.00	0.00
34	.74	1.30	0.00	0.00	0.00	0.00
35	.74	1.30	0.00	0.00	.35	.26
36	.74	1.30	.35	.26	.70	.52
37	.74	1.95	0.00	0.00	0.00	0.00
38	.74	1.95	.35	.26	.35	.26
39	1.00	−1.95	−.35	−.35	−.35	−.35
40	1.00	−1.95	0.00	0.00	0.00	0.00
41	1.00	−1.30	−.35	−.35	−.35	−.35
42	1.00	−1.30	0.00	0.00	0.00	0.00
43	1.00	−1.30	0.00	0.00	0.00	0.00
44	1.00	−1.30	.35	.35	.35	.35
45	1.00	−.65	−.35	−.35	−.70	−.70
46	1.00	−.65	−.35	−.35	−.35	−.35
47	1.00	−.65	0.00	0.00	−.35	−.35
48	1.00	−.65	0.00	0.00	0.00	0.00

continued on the next page

Table 2, continued
True Values of a, b_R, and d Parameters for Item Pools 1, 2, and 3
(d = 0.0 in Pool 1)

	Pools 1, 2, 3		Pool 2		Pool 3	
Item	a	b_R	d	ad	d	ad
49	1.00	–.65	.35	.35	0.00	0.00
50	1.00	–.65	.70	.70	.35	.35
51	1.00	0.00	–.70	–.70	–.35	–.35
52	1.00	0.00	–.35	–.35	–.35	–.35
53	1.00	0.00	–.35	–.35	0.00	0.00
54	1.00	0.00	0.00	0.00	0.00	0.00
55	1.00	0.00	0.00	0.00	0.00	0.00
56	1.00	0.00	.35	.35	.35	.35
57	1.00	0.00	.70	.70	.35	.35
58	1.00	.65	–.35	–.35	–.35	–.35
59	1.00	.65	–.35	–.35	–.35	–.35
60	1.00	.65	0.00	0.00	0.00	0.00
61	1.00	.65	0.00	0.00	0.00	0.00
62	1.00	.65	0.00	0.00	0.00	0.00
63	1.00	.65	.35	.35	.35	.35
64	1.00	.65	.35	.35	.35	.35
65	1.00	1.30	–.35	–.35	–.35	–.35
66	1.00	1.30	–.35	–.35	0.00	0.00
67	1.00	1.30	0.00	0.00	0.00	0.00
68	1.00	1.30	0.00	0.00	.35	.35
69	1.00	1.30	.35	.35	.35	.35
70	1.00	1.30	.35	.35	.70	.70
71	1.00	1.95	–.35	–.35	–.35	–.35
72	1.00	1.95	0.00	0.00	0.00	0.00
73	1.00	1.95	0.00	0.00	.35	.35
74	1.00	1.95	.35	.35	.35	.35
75	1.00	1.95	.35	.35	.70	.70

F group than it is for comparable members of the R group. An estimated standard error (SE) for $\text{MH}_{\text{D-DIF}}$, based on work by Robins, Breslow, & Greenland (1986) and Phillips & Holland (1987), is given in Holland & Thayer (1988) as

$$\text{SE}(\text{MH}_{\text{D-DIF}}) = 2.35\left\{\text{Var}\left[\ln(\hat{\alpha}_{\text{MH}})\right]\right\}^{1/2}, \tag{3}$$

where $\text{Var}\left[\ln(\hat{\alpha}_{\text{MH}})\right]$ is estimated by

$$\frac{\sum_k U_k V_k / T_k^2}{2\left(\sum_k A_k D_k / T_k\right)^2}, \tag{4}$$

where

$$U_k = (A_k D_k) + \hat{\alpha}_{\text{MH}}(B_k C_k) \tag{5}$$

and

$$V_k = (A_k + D_k) + \hat{\alpha}_{\text{MH}}(B_k + C_k). \tag{6}$$

The MH χ^2 test of the null hypothesis of no DIF was not examined here.

The matching variable for the DIF analysis of the CAT-administered items was obtained by (1) obtaining the simulee's MLE of θ, based on the responses to the 25 CAT items; and (2) using this MLE, along with the estimated item parameters, to compute an expected true score on all pool items by summing the 75 values of the estimated IRFs at $\hat{\theta}_{\text{CAT}}$. That is, the matching variable was

$$\text{Expected true score based on CAT} = \sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_{\text{CAT}}), \tag{7}$$

where $\hat{P}_j(\cdot)$ is an estimate of the 3PLM IRF, and $\hat{\theta}_{\text{CAT}}$ is the MLE of θ based on the CAT items. Examinees whose expected true scores fell in the same one-unit interval were considered matched.

Sample Size Conditions

It was not clear how best to define sample size for purposes of data simulation and analysis. If groups of a fixed sample size were drawn and the CAT administered, the sample sizes per item would have a very large range. For example, in Conditions 3 and 4 (see Table 1), the range of item sample sizes for the F group was from 0 to 51,133 (out of a total sample of 60,000). Because the goal was to investigate the behavior of DIF statistics for specific sample sizes, it would not have been useful to simply analyze the available data for each item. Therefore, several other approaches were considered.

Initially, an attempt was made to generate enough data to meet or exceed the target item sample sizes for all conditions. This required at least 900 R group members, as well as at least 500 members of each of the three F groups for each of the three pools (see Table 1). To achieve this goal for most items required generating 60,000 simulees for the R group and for each of the nine F distribution × item pool combinations. Even with 60,000 cases, five items with the lower value of a and medium or high values of b_{R} yielded a sample size of less than 500 for at least one pool-group combination. (This is in addition to the four items that were never administered.) To assess the variability of DIF results, two replications per condition were conducted.

Examination of the DIF results from this approach showed that the variability of the DIF statistics was large enough to prevent a useful characterization of the behavior of the statistics, even after averaging across two replications. Simulating additional CAT results was undesirable, however, because of the cost of data generation. Several resampling approaches that would have yielded multiple estimates of each statistic were considered, but none seemed ideal.

The approach that was ultimately selected was as follows: For each item, all the available CAT data (out of a maximum of 60,000 responses per group) were used to form the 2 (item response) $\times$ 2 (group) $\times K$ (level of the matching variable) contingency table needed for DIF analysis. The table frequencies then were converted to proportions of the total number of observations for the group in question. Using these proportions as estimates of the population probabilities associated with the $4 \times K$ cells for the relevant configuration of conditions, expected tables for the target sample sizes were obtained by multiplying the probability estimates for focal group cells by the desired F group sample size and then doing the same for the R group cells. Next, DIF statistics and SEs were computed, based on the expected tables, for all 18 conditions.

As a simple example of this expected table (ET) approach, consider the following hypothetical results for a single item, assuming that there are only two levels of the matching variable. First, use all available data for the item to construct a $2 \times 2 \times 2$ frequency table (because $K = 2$ here). Then divide the cell frequencies for the R group by the total number of R group examinees, and divide the frequencies for the F group by the total number of F group examinees. This produces a $2 \times 2 \times 2$ table of estimated probabilities (see Table 3).

Table 3
Probability and Target Values for the Expected Table Approach

Table and	Low on Matching Variable			High on Matching Variable		
Group	Correct	Incorrect	Total	Correct	Incorrect	Total
Probability Table						
Reference	.2	.1	.3	.5	.2	.7
Focal	.2	.2	.4	.4	.2	.6
Target Table						
Reference	180	90	270	450	180	630
Focal	20	20	40	40	20	60

Now suppose that target tables are needed for the $N_{\mathrm{R}} = 900, N_{\mathrm{F}} = 100$ condition. The R group probabilities are multiplied by 900 and the F group probabilities are multiplied by 100, producing a target table for use in DIF analysis (see Table 3).

$\mathrm{MH}^*_{\text{D-DIF}}$ and SE* denote DIF statistics computed from the expected tables with Equations 1–4. ($\mathrm{MH}_{\text{D-DIF}}$ and SE are used when the computations are not based on the expected tables or when these terms are used generically.) Note that SE* does not provide a measure of the error associated with the estimation of $\mathrm{MH}^*_{\text{D-DIF}}$. Instead, SE* closely approximates the SE that would be obtained using actual samples of the target sizes. The appropriate formula for the SE of the $\mathrm{MH}^*_{\text{D-DIF}}$ estimate can be derived as follows. Let $\hat{\alpha}^*_{\mathrm{MH}}$ be the ET-estimated MH odds ratio based on the target sample sizes N^*_{R} and N^*_{F} and $\mathrm{MH}^*_{\text{D-DIF}}$ be the corresponding DIF index. Then, based on the results in Equations 3 and 4,

$$\mathrm{SE}_{\mathrm{ET}} \equiv \mathrm{SE}_{\mathrm{ET}}\left(\mathrm{MH}^*_{\text{D-DIF}}\right) = 2.35\left\{\mathrm{Var}\left[\ln\left(\alpha^*_{\mathrm{MH}}\right)\right]\right\}^{1/2} \tag{8}$$

where $\mathrm{Var}\left[\ln\left(\alpha^*_{\mathrm{MH}}\right)\right]$ is estimated by

$$\frac{\sum_k U^*_k V^*_k / T^{*2}_k}{2\left(\sum_k A_k D_k / T^*_k\right)^2}, \tag{9}$$

where

$$U^*_k = (A_k D_k) + \hat{\alpha}^*_{\mathrm{MH}}(B_k C_k), \tag{10}$$

$$V^*_k = (A_k + D_k) + \hat{\alpha}^*_{\mathrm{MH}}(B_k + C_k), \tag{11}$$

and

$$T^*_k = \frac{N_{\mathrm{R}k} N^*_{\mathrm{R}}}{N_{\mathrm{R}}} + \frac{N_{\mathrm{F}k} N^*_{\mathrm{F}}}{N_{\mathrm{F}}}\ . \tag{12}$$

$\mathrm{SE}_{\mathrm{ET}}$ reflects the degree of precision with which the population DIF value is estimated using the ET approach. Because the ET estimate, $\mathrm{MH}^*_{\text{D-DIF}}$, usually was based on thousands of cases in this study, $\mathrm{SE}_{\mathrm{ET}}$ was typically much smaller than the ordinary SEs that would be obtained for the target sample sizes in question. In this study, the value of $\mathrm{SE}_{\mathrm{ET}}$ was found to be very similar to the value of SE (Equations 3–4) obtained using all the available data for the item.

Although it produces only a single estimate, the ET approach can provide a relatively precise idea of the behavior of $\text{MH}_{\text{D-DIF}}$. A supplementary study comparing the ET method to an estimation procedure based on multiple replications (as in a typical simulation study) appeared in Zwick, Thayer, & Wingersky (1993). The comparison was based on items for which 60,000 responses per population group were available. The ET method was found to give results similar to those of the replication-based approach. For the items that were studied, the ET estimate ($\text{MH}^{*}_{\text{D-DIF}}$) was determined to be as precise as an average over 316 replications of the $\text{MH}_{\text{D-DIF}}$ statistic based on the target sample sizes. Another advantage of the ET approach is that, once the $2 \times 2 \times K$ probability tables have been created, DIF results can be generated easily for any target sample size, facilitating further research.

Results

Comparison of CAT-Based and Nonadaptive DIF Analyses

For selected simulation conditions, MH results from the CAT analyses were compared to results of two nonadaptive DIF analyses. The first was a procedure (T-75) in which all 75 pool items were administered and simulees were matched on expected true score calculated using the MLE of θ based on all 75 responses (the nonadaptive control). That is, instead of the matching variable in Equation 7, the matching variable was

$$\text{Expected true score based on all 75 items} = \sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_{75}), \tag{13}$$

where $\hat{\theta}_{75}$ is the MLE of θ based on all 75 items. The second approach (NC) was a conventional DIF analysis, in which all 75 pool items were administered and simulees were matched on number-correct score. The results of this comparison are given in Tables 4 and 5.

This analysis included only the simulation conditions that had DIF and were based on R and F sample

Table 4
Correlations (Based on 71 Items) for True DIF (*ad*) and $\text{MH}_{\text{D-DIF}}$ Statistics Based on Three Types of Matching Variables (CAT, T-75, and NC) for Conditions 4, 10, and 16 (Pool 2), and Conditions 6, 12, and 18 (Pool 3)

Variables and	Condition						
Correlation	4	6	10	12	16	18	Median
CAT, T-75							
Uncorrected	.83	.88	.89	.88	.91	.89	.89
Corrected	.93	1.00[a]	1.00	.97	1.00[a]	.99	.99
CAT, NC							
Uncorrected	.85	.87	.89	.86	.90	.90	.88
Corrected	.96	.99	.99	.96	1.00	1.00[a]	.99
CAT, *ad*							
Uncorrected	.96	.95	.98	.96	.99	.96	.96
Corrected	.97	.96	.99	.97	1.00	.97	.97
T-75, NC							
Uncorrected	.99	.99	.99	.99	.99	.99	.99
Corrected	1.00[a]	1.00[a]	1.00[a]	1.00[a]	1.00[a]	1.00[a]	1.00[a]
T-75, *ad*							
Uncorrected	.84	.86	.88	.85	.90	.88	.87
Corrected	.93	.97	.98	.93	.99	.98	.97
NC, *ad*							
Uncorrected	.86	.87	.88	.84	.89	.89	.88
Corrected	.95	.98	.98	.92	.99	.98	.98

[a] Corrected value was greater than 1.0.

Table 5
Means and SDs of $\mathrm{MH}_{\text{D-DIF}}$ Statistics for Three Matching Variables for Conditions 4, 10, and 16 (Pool 2) and Conditions 6, 12, and 18 (Pool 3)

Matching Variable, Number of Items, and Statistic	Condition 4	6	10	12	16	18	Median
CAT, 71 Items							
Mean	0.00	.02	.02	.03	.01	.05	.02
SD	.96	.89	.99	.94	1.02	.96	.96
T-75, 71 Items							
Mean	–.02	.01	–.01	0.00	–.01	–.04	–.01
SD	.97	.90	.92	.96	1.02	.97	.97
NC, 71 Items							
Mean	–.02	–.02	–.02	–.02	–.02	–.08	–.02
SD	.97	.88	.93	.99	1.03	.97	.97
NC, 75 Items							
Mean	.01	.01	.01	0.00	.02	–.04	.01
SD	.96	.87	.93	.98	1.02	.97	.97

sizes of 500—that is, Conditions 4, 6, 10, 12, 16, and 18 (see Table 1). For each of the six conditions, the correlation matrix was computed for four variables: the three kinds of DIF statistics and the true DIF (*ad*) for the item. Each correlation matrix was based on the 71 items that were administered in the CATs.

The CAT-based DIF statistics used in this analysis were computed using the ET method; the two other statistics (T-75 and NC) were computed based on actual samples of 500 from the R and F groups. Therefore, for most items, the CAT statistics were much more precisely determined. To avoid giving a spuriously inflated impression of the performance of the CAT analyses, correlations that were corrected for unreliability were computed, using the following formula:

$$r^{c}_{XY} = \frac{r_{XY}}{\left[(r_{XX})(r_{YY})\right]^{1/2}}\,, \tag{14}$$

where r^{c}_{XY} is the corrected correlation between X and Y (in this case, DIF statistics), r_{XY} is the Pearson correlation between X and Y, and r_{XX} and r_{YY} are the reliabilities of X and Y. (If reliability is underestimated, the corrected correlation can exceed 1.) Reliability was estimated as

$$\text{Reliability} = 1 - \frac{\sum_{j} \mathrm{SE}^2_j\,(\mathrm{MH}_{\text{D-DIF}})/J}{\text{Variance across all } J \text{ items of } \mathrm{MH}_{\text{D-DIF}}}\,, \tag{15}$$

where J is the number of items. The numerator represents error variance, and the denominator represents total variance. [For the CAT DIF statistics, the $\mathrm{SE}^2_j(\cdot)$ values were the squares of the SE_{ET} values. The reliability of *ad* is 1.0 because it is not a statistic.] These corrected correlations provide a more equitable approach to comparing the CAT, T-75, and NC analyses.

The uncorrected and corrected intercorrelations of the three kinds of $\mathrm{MH}_{\text{D-DIF}}$ statistics and true DIF values are given in Table 4 for each of the six conditions. The median across conditions also is given. Table 4 shows that the CAT, T-75, and NC analyses produced results that were highly correlated with each other and with the true DIF values.

In particular, the results of the two analyses based on all 75 item responses (T-75 and NC) had near perfect correlations. (The similarity between these approaches may be substantially less for shorter tests.) The median corrected correlation with true DIF was approximately the same for the CAT, T-75, and NC analyses, which is somewhat surprising because the CAT DIF approach used θ estimates based on only 25 item responses.

A supplementary analysis based on a subset of the 18 simulation conditions showed that the ET method provided similar correlation results for the CAT-based DIF statistics as did an analysis based on actual samples of the target sizes. For example, for Condition 6, MH results based on actual samples of 500 per group were compared to the ET results. Based on the samples of 500, the uncorrected correlations of the CAT $\text{MH}_{\text{D-DIF}}$ statistics with $\text{MH}_{\text{D-DIF}}$ values from the T-75 and NC procedures were .88 and .87, respectively--the same as for the ET-based CAT statistics. Based on the samples of 500, the uncorrected correlation of the CAT $\text{MH}_{\text{D-DIF}}$ statistics with true DIF was .92, compared to .95 for the ET method.

Table 5 shows the mean and SD of the $\text{MH}_{\text{D-DIF}}$ values for the CAT, T-75, and NC analyses across the 71 items given in the CAT. Results are shown for each simulation condition, along with the medians over the six simulation conditions. The mean across 71 items of the *ad* values was .007 in Conditions 4, 10, and 16 (Pool 2) and −.001 in Conditions 6, 12, and 18 (Pool 3). The SDs were .288 and .293 for Pools 2 and 3, respectively. Because $\text{MH}_{\text{D-DIF}}$ is approximately equal to $3ad$, $\text{MH}_{\text{D-DIF}}$ would be expected to have a mean of approximately 0.0 and a SD of approximately .90.

In DIF analyses in which all examinees take all items and the matching variable is number-correct score, the average $\text{MH}_{\text{D-DIF}}$ is constrained to be approximately 0.0 across items, producing a negative covariance among the DIF statistics within a test. This constraint on the $\text{MH}_{\text{D-DIF}}$ was not present in the CAT and T-75 analyses. In these types of DIF analysis, the nature of the covariance across DIF statistics within a test is unknown.

The issue of covariances across DIF statistics is relevant to Table 5 for two reasons. First, because of the constraint on the mean of the NC-based statistics, it was not clear which across-item NC mean was the most useful for comparison to other analyses--the one based on only the 71 items given in the CAT or the mean over 75 items. Both these means and the corresponding SDs are, therefore, included in Table 5. Second, the nonzero covariances for the NC-based statistics and possibly for the other DIF statistics made it difficult to estimate the SEs of the means in Table 5. If the $\text{MH}_{\text{D-DIF}}$ statistics were independent across items, the SEs of the average $\text{MH}_{\text{D-DIF}}$ statistics in Table 5 would be approximately .009 for the CAT analysis and .049 for the two nonadaptive analyses (obtained by dividing the average item-level SE by the square root of the number of items). Judged in this light, the means for the nonadaptive procedures were quite close to 0.0, but the means for the CAT procedure were slightly inflated. All six means for the CAT-based procedure were greater than 0.0 and the means were larger for the Pool 3 conditions (.02, .03, .05) than for the Pool 2 conditions (0.00, .02, .01).

However, these values for the SE of the mean are only approximate. Because of the negative covariances among NC $\text{MH}_{\text{D-DIF}}$ statistics within a test, the value of .049 is an overestimate of the SE of the mean for the NC analyses. Presumably, this overestimation holds for the T-75 approach, which produced results nearly identical to the NC analyses. For the CAT DIF statistics, the value computed under independence may either underestimate or overestimate the SE. In any case, the practical implications of an inflation of .01 to .05 in the $\text{MH}_{\text{D-DIF}}$ statistic are small in that a difference this size is unlikely to have much effect on decisions about the item. It would be possible, of course, to rescale the statistics so that they would be centered on 0.0 for a particular collection of items.

The SEs of $\text{MH}_{\text{D-DIF}}$ also were compared across the three matching variables. For the T-75 and NC analyses, the average values of $\text{SE}(\text{MH}_{\text{D-DIF}})$ within each condition were approximately .40, whereas the CAT-based $\text{MH}^*_{\text{D-DIF}}$ statistics tended to have SEs of approximately .35. One hypothesis for this discrepancy is that the smaller SEs for the CAT DIF analysis were related to the use of the ET estimation method. The supplementary study of the ET method did show small differences due to estimation method. The ET method produced SE estimates (SE*) that were, on the average, .02 less than the empirical SD of $\text{MH}_{\text{D-DIF}}$ across replications. The ordinary estimation method produced SEs that were, on the average, .01 greater than the empirical SD. Based on a follow-up study, it appears that the smaller SEs in CAT-based DIF analysis resulted primarily from the restriction of the analysis to examinees in a smaller θ range, rather than from the use of the ET method (Zwick, Thayer, & Wingersky, 1994).

$\text{MH}^*_{\text{D-DIF}}$ Statistics for Combinations of *ad* and b_R

The average CAT DIF statistics for various configurations of item parameters and simulation factors were examined. To determine the best way to summarize the results, a series of analyses of variance (ANOVAs) was conducted in which the observations were the DIF statistics and the independent variables were sample size condition, F group distribution, item difficulty level, item discrimination, item DIF level, and item position. Pool 1 was analyzed separately; Pools 2 and 3 were analyzed both separately and in combination (with pool as an additional independent variable). The $\text{MH}^*_{\text{D-DIF}}$ statistics were analyzed under several different assumptions concerning interactions among the independent variables and several different numbers of levels of item difficulty, item position, and DIF.

Results were quite consistent across the analysis models. In Pool 1, only the item difficulty effect was significant at an α of .01; it explained less than 3% of the variance in the $\text{MH}^*_{\text{D-DIF}}$ statistics. (As in all exploratory analyses, significance testing can be viewed here only as a rough tool for ranking the size of effects.) In Pools 2 and 3, the level of DIF explained approximately 85% of the variance. Most analyses of Pools 2 and 3 showed very small, but statistically significant effects of item difficulty level, and of the item difficulty × DIF level and item discrimination × DIF level interactions. These results are consistent with those of Donoghue et al. (1993). Sample size had no effect. (Because the results for the two sample size conditions were generated from the same set of expected tables, they were highly correlated. The main value of generating results for two sample size conditions was that it allowed the examination of the behavior of the SEs of the DIF statistics, discussed below.) Item position and pool never yielded statistically significant main effects, but these factors sometimes showed very small interactions with item difficulty and discrimination.

The effects of F group distribution and its interactions with other variables were minimal. In Pool 1, the average values of $\text{MH}^*_{\text{D-DIF}}$ were ordered in the same way as the F groups, but the differences across groups were small: The averages were −.008, .003, and .011 for analyses involving the $N(-1,1)$, $N(0,1)$, and $N(.5,1)$ focal groups, respectively. The same ordering was evident for the no-DIF items in Pools 2 and 3. For items that did have DIF, the values of $\text{MH}^*_{\text{D-DIF}}$ tended to be closer to 0.0 for analyses involving the $N(-1,1)$, focal group than for analyses based on the other two focal groups; but, again, the effect was very small.

In general, these findings are fairly consistent with previous studies. For example, the simulation study of Shealy & Stout (1993) showed that the difference in location between reference and focal groups had little effect on the power and Type I error rate of the MH χ^2 statistic, although location differences did have a small effect on the values of $\text{MH}_{\text{D-DIF}}$ in the non-null case.

Based on the ANOVA findings, $\text{MH}^*_{\text{D-DIF}}$ means were examined for every combination of *ad* and b_R. The average $\text{MH}^*_{\text{D-DIF}}$ for Pool 1, where $ad = 0$, was 0.0 for all values of b_R. The average $\text{MH}^*_{\text{D-DIF}}$ statistics are given in Table 6 for Pools 2 and 3 for the $N_R = 500$, $N_F = 500$ condition. As noted, $\text{MH}^*_{\text{D-DIF}}$ results were nearly identical for the two sample size conditions. The average SE of the estimate, SE_{ET}, is provided as well. The average value of SE_{ET} is the maximum value that the SE of the mean $\text{MH}^*_{\text{D-DIF}}$ could take (i.e., the value that would occur if all items had intercorrelations of 1.0) and therefore is an overestimate of the SE of the mean. Because results were averaged over the three F group distributions, a single item within a pool generated three values.

Table 6
Average $MH^*_{D\text{-}DIF}$, Average $SE_{ET}(MH^*_{D\text{-}DIF})$, and the Number of Items Over Which the Averages Were Computed for Each Combination of *ad* and b_R for $N_R = 500$ and $N_F = 500$ for Pools 2 and 3

Pool, b_R, Statistic, and Number of Items	Value of *ad*									
	−.70	−.52	−.35	−.26	0	.26	.35	.52	.70	Average
Pool 2, b_R = −1.95										
$MH^*_{D\text{-}DIF}$			−1.3	−.9	0.0	1.0				−.3
SE_{ET}			.07	.08	.10	.10				.09
n Items			3	6	9	3				21
Pool 2, b_R = −1.30										
$MH^*_{D\text{-}DIF}$			−1.3	−.9	0.0	.9	1.3			0.0
SE_{ET}			.05	.08	.06	.09	.06			.07
n Items			3	6	12	6	3			30
Pool 2, b_R = −.65										
$MH^*_{D\text{-}DIF}$			−1.3	−.9	0.0	.8	1.2		2.4	0.0
SE_{ET}			.04	.08	.05	.10	.04		.05	.06
n Items			6	6	15	6	3		3	39
Pool 2, b_R = 0.0										
$MH^*_{D\text{-}DIF}$	−2.1	−2.2	−1.2	.1	0.0	.9	1.2	1.8	2.5	.1
SE_{ET}	.04	.30[a]	.04	.58[a]	.04	.05	.04	.08	.05	.11
n Items	3	3	6	3	9	6	3	3	3	39
Pool 2, b_R = .65										
$MH^*_{D\text{-}DIF}$			−1.2	−.9	0.0	1.0	1.2			0.0
SE_{ET}			.06	.06	.08	.05	.04			.07
n Items			6	3	15	3	6			33
Pool 2, b_R = 1.30										
$MH^*_{D\text{-}DIF}$		−1.8	−.9	−.5	0.0	1.0	1.1			−.1
SE_{ET}		.10	.06	.16	.10	.11	.06			.09
n Items		3	6	3	12	3	6			33
Pool 2, b_R = 1.95										
$MH^*_{D\text{-}DIF}$			−.7		0.0	1.1	.8			.3
SE_{ET}			.09		.09	.20	.06			.10
n Items			3		6	3	6			18
Pool 2, Average										
$MH^*_{D\text{-}DIF}$	−2.1	−2.0	−1.1	−.7	0.0	.9	1.1	1.8	2.5	0.0
SE_{ET}	.04	.20	.05	.14	.07	.09	.05	.08	.05	.08
n Items	3	6	33	27	78	30	27	3	6	213
Pool 3, b_R = −1.95										
$MH^*_{D\text{-}DIF}$		−1.6	−.9	−.7	.3	1.3				−.3
SE_{ET}		.08	.07	.09	.09	.09				.09
n Items		3	3	6	6	3				21
Pool 3, b_R = −1.30										
$MH^*_{D\text{-}DIF}$		−1.4	−.9	−.6	.3	1.2	1.6			0.0
SE_{ET}		.07	.05	.07	.06	.08	.06			.06
n Items		3	3	6	12	3	3			30
Pool 3, b_R = −.65										
$MH^*_{D\text{-}DIF}$	−2.2		−1.0	−.7	.2	1.0	1.5			−.1
SE_{ET}	.04		.04	.08	.05	.09	.04			.06
n Items	3		6	6	15	6	3			39

continued on the next page

Table 6, continued

Average $\mathrm{MH}^*_{\text{D-DIF}}$, Average $\mathrm{SE}_{\mathrm{ET}}(\mathrm{MH}^*_{\text{D-DIF}})$, and the Number of Items Over Which the Averages Were Computed for Each Combination of *ad* and b_R for $N_R = 500$ and $N_F = 500$ for Pools 2 and 3

Pool, b_R, Statistic, and Number	Value of *ad*									
of Items	−.70	−.52	−.35	−.26	0	.26	.35	.52	.70	Average
Pool 3, b_R = 0.0										
$\mathrm{MH}^*_{\text{D-DIF}}$			−1.1	−.6	.1	1.0	1.3			.1
$\mathrm{SE}_{\mathrm{ET}}$			.04	.46[a]	.04	.06	.04			.11
n Items			6	6	15	6	6			39
Pool 3, b_R = .65										
$\mathrm{MH}^*_{\text{D-DIF}}$			−1.4	−1.0	−.1	.8	1.1			−.1
$\mathrm{SE}_{\mathrm{ET}}$			.06	.06	.08	.05	.05			.07
n Items			6	3	15	3	6			33
Pool 3, b_R = 1.30										
$\mathrm{MH}^*_{\text{D-DIF}}$			−1.1	−1.3	−.3	.6	.9	1.6	2.1	.2
$\mathrm{SE}_{\mathrm{ET}}$			.05	.09	.11	.08	.05	.11	.08	.09
n Items			3	3	12	3	6	3	3	33
Pool 3, b_R = 1.95										
$\mathrm{MH}^*_{\text{D-DIF}}$			−1.1		−.5	.6	.6		1.5	.3
$\mathrm{SE}_{\mathrm{ET}}$			.08		.11	.17	.06		.06	.09
n Items			3		3	3	6		3	18
Pool 3, Average										
$\mathrm{MH}^*_{\text{D-DIF}}$	−2.2	−1.5	−1.1	−.7	0.0	.9	1.1	1.6	1.8	0.0
$\mathrm{SE}_{\mathrm{ET}}$	.04	.08	.05	.15	.07	.09	.05	.11	.07	.08
n Items	3	6	30	30	78	27	30	3	6	213

[a] The average $\mathrm{SE}_{\mathrm{ET}}(\mathrm{MH}^*_{\text{D-DIF}})$ was large because of the sparsity of data for one item.

Table 6 shows that the average value of $\mathrm{MH}^*_{\text{D-DIF}}$ was typically about 3.3 times the value of *ad* in Pool 2, and 3 times the value of *ad* in Pool 3. In Pool 2, for a fixed value of *ad*, the average $\mathrm{MH}^*_{\text{D-DIF}}$ usually decreased in absolute value as b_R increased. For example, for $ad = -.35$, the average $\mathrm{MH}^*_{\text{D-DIF}}$ was -1.3 for $b_R = -1.95$, -1.2 for $b_R = 0.0$, and $-.7$ for $b_R = 1.95$. This phenomenon, noted by Donoghue et al. (1993), occurs in simulations in which the guessing parameter *c* is constrained to be the same in the reference and focal groups. The more difficult the item, the closer the probability of correct response is to the guessing value, and the more difficult the groups are to differentiate. Superimposed on this phenomenon, Pool 3 (Table 6) included a correlation between the b_R and DIF parameters. Easier items in Pool 3 were more likely to have negative DIF than more difficult items. The relation between $\mathrm{MH}^*_{\text{D-DIF}}$ and b_R for fixed *ad* was not as evident in Pool 3 as it was in Pool 2. Also, the average $\mathrm{MH}^*_{\text{D-DIF}}$ for the no-DIF items were not as close to 0.0 as they were in Pools 1 and 2. For $d = 0$, the average $\mathrm{MH}^*_{\text{D-DIF}}$ decreased from .3 to −.5 as b_R increased from −1.95 to 1.95.

The values of SE* (not shown) varied little across pools, DIF levels, b_R, or *a*. The primary determinant of SE* was sample size. For the $N_R = 500, N_F = 500$ conditions, SE* ranged from approximately .3 to .4; for the $N_F = 900, N_F = 100$ conditions, the range was from approximately .5 to .7. Using the estimated value of the SE for a particular pair of sample sizes (e.g., N_{R1}, N_{F1}) as a baseline, the SE of $\mathrm{MH}_{\text{D-DIF}}$ for another sample size pair (e.g., N_{R2}, N_{F2}) can be predicted accurately using the ratio of the harmonic means of the sample size pairs (Zwick et al., 1994). More specifically, SE(N_{R2}, N_{F2}) can be well predicted by multiplying SE(N_{R1}, N_{F1}) by

$$\left[\frac{h(N_{R1},N_{F1})}{h(N_{R2},N_{F2})}\right]^{1/2} \qquad (16)$$

where $h(\cdot)$ denotes the harmonic mean.

Estimated Percent of "C" Results for Combinations of *ad* and b_R

ETS has a system for categorizing the severity of DIF based on MH results. According to this classification scheme, a "C" categorization, which represents large DIF, requires that the absolute value of $\text{MH}_{\text{D-DIF}}$ be at least 1.5 and be significantly greater than 1 (at α =.05). A "B" categorization, which indicates moderate DIF, requires that $\text{MH}_{\text{D-DIF}}$ be significantly different from 0.0 (at α =.05) and that the absolute value of $\text{MH}_{\text{D-DIF}}$ be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to be free of DIF. Items that fall in the C category are typically eliminated from tests or subjected to further scrutiny.

Because most of the ET estimates, $\text{MH}^*_{\text{D-DIF}}$ and SE*, were based on at least 10,000 observations, it is reasonable to assume that they provided precise estimates of the population mean and SD of $\text{MH}_{\text{D-DIF}}$ for the relevant configuration of item properties and simulation conditions. This is supported by the supplementary analysis reported in Zwick, Thayer, & Wingersky (1993). If it is assumed that $\text{MH}_{\text{D-DIF}}$ statistics for this configuration follow a normal distribution with this mean and SD, percentiles of the theoretical distribution of $\text{MH}_{\text{D-DIF}}$ can be obtained. These percentiles then can be used to estimate the percent of times such an item would be classified as an A, B, or C item in repeated administrations of the test. This is an alternative way of providing information about the sampling variation of the $\text{MH}_{\text{D-DIF}}$ statistic. Viewing an item's DIF status as probabilistic, rather than deterministic, may be a fruitful way of evaluating DIF results.

Based on the ETS DIF rules, an algorithm was developed for estimating these percents, to be applied separately to each item in each condition. The algorithm was tested and found to work well with data for 15 items from the simulation, using the ET estimates, $\text{MH}^*_{\text{D-DIF}}$and SE*, to approximate the mean and SD of the $\text{MH}_{\text{D-DIF}}$ distribution. The algorithm also worked well on data from the simulation study of Donoghue et al. (1993), using the average over 100 replications of $\text{MH}_{\text{D-DIF}}$ and SE to estimate the mean and SD of the $\text{MH}_{\text{D-DIF}}$ distribution. Details are given in Zwick, Thayer, & Wingersky (1993).

Because of the complexity of the relation between *ad* and $\text{MH}_{\text{D-DIF}}$ in the 3PLM, the determination of which items are nominally A, B, and C items is not straightforward. Based on the empirical finding that $\text{MH}_{\text{D-DIF}}$ was approximately equal to 3*ad* in the conditions investigated in this study, items with $ad = \pm.70$ and $ad = \pm.52$ were considered to be nominal C items, those with $ad = \pm.35$ to be nominal B items, and those with $ad = \pm.26$ or $ad = 0$ to be nominal A items. (If, instead, the nominal categorization were based on the Rasch model finding that $\text{MH}_{\text{D-DIF}}$ provides an estimate of 4*ad*, the only change would be that items with $ad = \pm.26$ would be considered nominal B items.)

For Pool 1, the average expected percents of C results were 0.0 for all values of b_R in the $N_R = 900, N_F = 100$ condition; in the $N_R = 500, N_F = 500$ condition, the percents were .1 for all b_R levels except for $b_R = -1.95$, in which the percent was .2. Table 7 gives the average expected percent of C results for each combination of *ad* and b_R for Pools 2 and 3. Like Table 6, Table 7 is averaged over the three focal group distributions.

Table 7 shows that with samples of 500 in each group, Pool 2 items with $ad = \pm.70$ would nearly always be identified as C items. Those with $ad = \pm.52$ would be expected to be so labeled at least three quarters of the time. As anticipated, the power to detect extreme DIF items was substantially smaller for the $N_R = 900, N_F = 100$ sample size condition. Table 7 shows that detection rates for the nominal C items in Pool 3 were smaller than in Pool 2.

Table 7
Average Expected Percent of C for $N_R = 900, N_F = 100$ (AEP),
Average Expected Percent for $N_R = 500, N_F = 500$ (AEP500), and
Number of Items for Each Combination of *ad* and b_R for Pools 2 and 3

Pool, b_R, Statistic, and Number	Value of *ad*									
of Items	−.70	−.52	−.35	−.26	0	.26	.35	.52	.70	Average
Pool 2, $b_R = -1.95$										
AEP			10.3	3.6	.2	4.2				3.2
AEP500			15.1	3.3	0.0	3.9				3.6
n Items			3	6	9	3				21
Pool 2, $b_R = -1.30$										
AEP			11.9	3.4	.1	3.4	11.1			3.7
AEP500			19.5	2.9	0.0	2.7	18.7			4.9
n Items			3	6	12	6	3			30
Pool 2, $b_R = -.65$										
AEP			11.4	3.5	.1	2.2	9.8		67.2	8.6
AEP500			18.7	3.2	0.0	1.4	15.5		97.3	12.3
n Items			6	6	15	6	3		3	39
Pool 2, $b_R = 0.0$										
AEP	60.5	49.6	8.2	.5	.1	3.7	7.8	38.3	78.3	19.9
AEP500	93.9	83.0	11.6	.1	0.0	3.1	11.6	75.3	99.5	30.2
n Items	3	3	6	3	9	6	3	3	3	39
Pool 2, $b_R = .65$										
AEP			8.5	2.7	.1	4.3	8.7			3.8
AEP500			12.0	1.8	0.0	4.0	12.8			5.0
n Items			6	3	15	3	6			33
Pool 2, $b_R = 1.30$										
AEP		44.8	4.1	.8	.1	5.2	8.1			6.9
AEP500		81.4	4.0	.2	0.0	5.7	11.5			10.8
n Items		3	6	3	12	3	6			33
Pool 2, $b_R = 1.95$										
AEP			2.0		.1	7.5	3.3			2.7
AEP500			1.2		0.0	11.3	2.8			3.0
n Items			3		6	3	6			18
Pool 2, Average										
AEP	60.5	47.2	8.1	2.8	.1	4.0	7.7	38.3	72.7	7.9
AEP500	93.9	82.2	11.7	2.3	0.0	3.9	11.1	75.3	98.4	11.5
n Items	3	6	33	27	78	30	27	3	6	213
Pool 3, $b_R = -1.95$										
AEP		26.8	3.2	1.2	.4	12.5				6.6
AEP500		52.1	2.5	.5	.1	21.9				11.1
n Items		3	3	6	6	3				21
Pool 3, $b_R = -1.30$										
AEP		17.6	3.6	.9	.3	10.0	22.0			5.6
AEP500		34.1	3.1	.2	0.0	15.3	45.0			9.8
n Items		3	3	6	12	3	3			30
Pool 3, $b_R = -.65$										
AEP	63.8		4.2	1.7	.1	5.5	18.8			8.2
AEP500	95.2		4.0	.9	0.0	6.2	37.8			11.9
n Items	3		6	6	15	6	3			39

continued on the next page

Table 7, continued
Average Expected Percent of C for $N_R = 900, N_F = 100$ (AEP),
Average Expected Percent for $N_R = 500, N_F = 500$ (AEP500), and
Number of Items for Each Combination of *ad* and b_R for Pools 2 and 3

Pool, b_R, Statistic, and Number	Value of *ad*									
of Items	−.70	−.52	−.35	−.26	0	.26	.35	.52	.70	Average
Pool 3, b_R = 0.0										
AEP			6.4	2.6	.1	5.0	14.2			4.4
AEP500			8.0	2.6	0.0	5.3	26.0			6.4
n Items			6	6	15	6	6			39
Pool 3, b_R = .65										
AEP			15.6	4.6	.1	2.3	7.0			4.8
AEP500			27.7	4.4	0.0	1.2	9.1			7.2
n Items			6	3	15	3	6			33
Pool 3, b_R = 1.30										
AEP			8.6	12.1	.3	.8	3.6	26.3	54.7	10.1
AEP500			12.0	19.4	.1	.2	3.2	55.1	92.4	16.9
n Items			3	3	12	3	6	3	3	33
Pool 3, b_R = 1.95										
AEP			8.4		.7	1.2	1.0		27.1	6.6
AEP500			11.8		.1	.4	.3		48.1	10.2
n Items			3		3	3	6		3	18
Pool 3, Average										
AEP	63.8	22.2	7.6	3.0	.2	5.3	9.3	26.3	40.9	6.6
AEP500	95.2	43.1	10.9	3.2	0.0	6.9	16.0	55.1	70.3	10.4
n Items	3	6	30	30	78	27	30	3	6	213

Summarizing results for the $N_R = 500, N_F = 500$ condition across Pools 2 and 3 (and across the possible definitions of the nominal categories), it would be expected that the nominal A items would be declared C items approximately 1% of the time using the investigated DIF procedures, the nominal B items would be identified as C items approximately 10% of the time, and the nominal C items would be correctly identified approximately 76% of the time. For the $N_R = 900, N_F = 100$ conditions, the corresponding percents were 1%, 7%, and 46% for the nominal A, B, and C items, respectively.

Discussion and Conclusions

The findings, in general, appear to be useful for testing programs that wish to establish DIF screening procedures for adaptively administered items. The CAT-based DIF statistics were found to be highly correlated with true DIF and with DIF measures based on nonadaptive administration. The mean DIF statistics for each pool were close to their nominal value of 0.0, although the CAT-based statistics showed a slight inflation, particularly for Pool 3, in which DIF and difficulty were positively correlated. Further analyses demonstrated that the modified DIF procedures would lead to reasonably accurate classification of items into three categories of DIF severity that are used by some testing programs in assembling forms. The detection rate for nominal C items was somewhat lower in Pool 3 than in Pool 2.

It is difficult to evaluate the importance of the finding that DIF statistics behaved somewhat better in Pool 2 than in Pool 3. Pool 3 was created because of the finding that DIF estimates are sometimes positively correlated with item difficulty estimates. This does not imply that the appropriate data-generating model is one in which the true (and ordinarily unknown) DIF and difficulty parameters are correlated. Thus, there is no compelling evidence for determining which of these pools is more realistic.

The factors that affected the size of the DIF estimates, in general, were the size of the true DIF, the item difficulty, and the interactions of true DIF with item difficulty and item discrimination. Focal group distribution, item position, and sample size had almost no effect. None of the results suggested that CAT-based DIF statistics differ in any substantial or consistent way from DIF statistics based on nonadaptive administration.

A finding that was useful, although not directly relevant to CATs, was that in nonadaptive administration of 75 items to 500 reference and 500 focal group examinees, matching on the expected true score based on the MLE of θ led to essentially the same results as matching on number-correct score. The use of matching variables based on item response theory in DIF analyses of both adaptive and nonadaptive tests is supported by the high correlations of the resulting DIF indexes with true DIF.

There are many questions that this study did not address. For example, the problem of insufficient item data, which may arise when conducting DIF analyses of adaptive tests (Miller, 1992), was not investigated. The effect of using alternative procedures, such as Bayesian methods, for estimating θs or item parameters was not examined. CAT algorithms that include item format and content constraints were not evaluated, nor were complex starting algorithms intended to control the exposure of items. Methods for refining the DIF criterion by deleting DIF items and repeating the analysis were not considered. These are all fruitful areas for future research.

References

Birnbaum, A. (1968). Some latent trait models. In F. Lord & M. Novick, *Statistical theories of mental test scores* (pp. 397–424). Reading MA: Addison-Wesley.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (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.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Research Rep. No. RR 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 & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty* (Rep. No. 89-5). New York: College Board.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11,* 23–27.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95–100.

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.

Miller, T. R. (1992, April). *Practical considerations for conducting studies of differential item functioning (DIF) in a CAT environment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Powers, D. E., & O'Neill, K. (1992). *Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills* (ETS Research Rep. No. RR 92-75). Princeton NJ: Educational Testing Service.

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.

Schaeffer, G., Reese, C., Steffen, M., McKinley, R., & Mills, C. N. (1993). *Field test of a computer-based GRE general test* (ETS Research Rep. No. RR 93-07). Princeton NJ: Educational Testing Service.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 187–231). Hillsdale NJ: Erlbaum.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST user's guide: LOGIST version 6.00.* Princeton NJ: 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., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., Thayer, D., & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests* (ETS Research Rep. No. RR 93-11). Princeton NJ: Educational Testing Service.

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

Acknowledgments

The authors thank several important contributors to this work. Paul Holland developed the initial version of the proposal that served as a guideline for the research. Charles Lewis provided excellent and speedy advice on several troublesome statistical issues. He proposed the "expected table" procedure used to estimate DIF statistics, including the SE formula in Equations 8 and 9. He also suggested the probabilistic approach used to report item DIF classifications. Neil Dorans, Nancy Petersen, and Ming Mei Wang made many thoughtful recommendations about the simulation and the report. Martha Stocking wrote the initial version of the CAT simulation program that was revised for the study and provided useful discussions about adaptive tests. Liz Brophy and Martha Byelich prepared the manuscript.

Author's Address

Send requests for reprints or further information to Rebecca Zwick, Educational Testing Service, Rosedale Road, Princeton NJ 08541, U.S.A. Internet: rzwick@ets.org.