DOCUMENT RESUME

ED 385 548 TM 023 964

AUTHOR	Zwick, Rebecca; And Others
TITLE	A Simulation Study of Methods for Assessing Differential Item Functioning in Computer-Adaptive Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-11
PUB DATE	Feb 93
NOTE	121p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Correlation; Error of Measurement; *Estimation (Mathematics); *Item Bias; Item Response Theory; Pretests Posttests; Simulation; *Test Items
IDENTIFIERS	*Mantel Haenszel Procedure; *Standardization; Three Parameter Model

ABSTRACT

Simulated data were used to investigate the performance of modified versions of the Mantel-Haenszel and standardization methods of differential item functioning (DIF) analysis in computer-adaptive tests (CATs). Each "examinee" received 25 items out of a 75-item pool. A three-parameter logistic item response model was assumed, and examinees were matched on expected true scores based on their CAT responses and on estimated item parameters. Both DIF methods performed well. The CAT-based DIF statistics were highly correlated with DIF statistics based on nonadaptive administration of all 75 pool items and with the true magnitudes of DIF in the simulation. DIF methods were also investigated for "pretest items," for which item parameter estimates were assumed to be unavailable. The pretest DIF statistics were generally well-behaved and also had high correlations with the true DIF. The pretest DIF measures, however, tended to be slightly smaller in magnitude than their CAT-based counterparts. Also, in the case of the Mantel-Haenszel approach, the pretest DIF statistics tended to have somewhat larger standard errors than the CAT-DIF statistics. Appendix A contains 10 supplementary tables; and Appendixes B, C, and D present additional information about the expected table estimator. Twenty-two tables in Appendix D present analysis results. (Contains 24 references.) (Author/SLD)

RESEARCH REPORT

A SIMULATION STUDY OF METHODS FOR ASSESSING DIFFERENTIAL ITEM FUNCTIONING IN COMPUTER-ADAPTIVE TESTS

Rebecca Zwick
Dorothy T. Thayer
Marilyn Wingersky

Educational Testing Service
Princeton, New Jersey
February 1993

February 10, 1993

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

Rebecca Zwick, Dorothy T. Thayer, and Marilyn Wingersky

Educational Testing Service

Table of Contents

Acknowledgments

We would like to thank several important contributors to this work. Paul Holland developed the initial version of the Holland and Zwick (1991) proposal that served as a guideline for our research. Charlie Lewis provided excellent and speedy advice on several troublesome statistical issues and suggested both the expected table approach and the estimation of the proportion of times each item would be placed into each of the three Educational Testing Service DIF categories. Our advisory committee, consisting of Neil Dorans, Nancy Petersen and Ming Mei Wang, provided thoughtful suggestions that helped us through some difficult decisions in planning and executing our simulation. Martha Stocking also provided useful discussions about adaptive tests, as well as a helpful review of our paper. Preliminary findings from this research were presented in Zwick et al. (1992) and Zwick (1992).

Abstract

Simulated data were used to investigate the performance of modified versions of the Mantel-Haenszel and standardization methods of differential item functioning (DIF) analysis in computer-adaptive tests (CATs). Each "examinee" received 25 items out of a 75-item pool. A three-parameter logistic item response model was assumed, and examinees were matched on expected true scores based on their CAT responses and on estimated item parameters. Both DIF methods performed well. The CAT-based DIF statistics were highly correlated with DIF statistics based on nonadaptive administration of all 75 pool items and with the true magnitudes of DIF in the simulation. DIF methods were also investigated for "pretest items," for which item parameter estimates were assumed to be unavailable. The pretest DIF statistics were generally well-behaved and also had high correlations with the true DIF. The pretest DIF measures, however, tended to be slightly smaller in magnitude than their CAT-based counterparts. Also, in the case of the Mantel-Haenszel approach, the pretest DIF statistics tended to have somewhat larger standard errors than the CAT DIF statistics.

1. Overview

Many large-scale testing programs are now developing or piloting computer-adaptive tests (CATs). Among these are the Scholastic Aptitude Test (SAT), the Graduate Record Examinations (GRE), and Praxis (successor to the NTE teacher assessment), developed at Educational Testing Service (ETS), the COMPASS placement tests produced by the American College Testing Program, the College Board Computerized Placement Tests, 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 that new approaches be developed for assessing validity and reliability and for analyzing item properties, including differential item functioning (DIF).

The purpose of our project was to investigate whether existing DIF analysis methods could be modified to accommodate the data collected in a CAT. There are several 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, item difficulty and DIF have been found to be positively related to an appreciable degree for some pairs of populations (e.g., Kulick & Hu, 1989). Therefore, if the group of primary interest--the *focal*

group--scores substantially below the comparison, or *reference* group, the CATs encountered by the focal group members will be made up of easier items than the CATs encountered by reference group members. If easier items have, on average, more negative DIF (i.e., DIF disadvantaging the focal group) than harder items, then the scores of focal group members may be lower than they should be and even lower than they would be on a comparable nonadaptive version of the test (Holland & Zwick, 1991). Finally, 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. Legg and Buhr (1992) and Schaeffer, Reese, and Steffen (1992) both report ethnic and gender group differences in some of these attributes. Their findings suggest that attitudes toward computer testing may be surprisingly complex. For example, Schaeffer, Reese, and Steffen (1992) found that Asian test-takers were most likely to have a computer available at home and most likely to report that using the computer mouse was very easy. Yet both Schaeffer et al. and Legg and Buhr found that Asian examinees were more likely than any other ethnic group to state that they preferred paper-and-pencil to computerized administration.

To investigate DIF detection in CATs, we simulated data consisting of responses to three different pools of 75 items. 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. The only kind of DIF that was studied 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 was the DIF status of the items and the item difficulties and discriminations.

Using a CAT algorithm based on item information, each "examinee" was assigned 25 items from one of the three pools of 75 items. Responses to the selected items were generated using the three-parameter logistic (3PL) item response theory model. The maximum likelihood estimate (MLE) of the examinee's ability was recomputed after each item was administered and the next item selected was the most informative item at the examinee's estimated ability.

The simulated data were used to investigate the feasibility of conducting DIF analyses using modified versions of the Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) and the standardization method of Dorans and Kulick (1986). Examinees were matched on the expected true score for the entire 75-item pool, computed using estimated ability from the 25 CAT items and estimated item parameters. An approach of this kind was suggested by Steinberg, Thissen, and Wainer (1990).

In addition, DIF analyses were conducted for "pretest" items that were administered nonadaptively. All examinees received the same set of pretest items, along with the CAT. For DIF analyses of the pretest items, the matching variable was the sum of the expected true score based on the CAT responses and the score (0 or 1) on the item under analysis, referred to as the studied item.

To disentangle the effects of assigning items via the CAT algorithm on one hand and matching examinees on expected true score on the other, we also included, for some simulation conditions, a "nonadaptive control" analysis in which the matching variable for

DIF analysis was the expected true score computed with the MLE estimated from 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-right score, as in conventional MH and standardization analysis.

The CAT-based DIF statistics were found to be highly correlated with true DIF and with DIF measures based on nonadaptive administration. Furthermore, the mean DIF statistics for each pool were close to their nominal value of zero. Although Pool 3 DIF statistics were not quite as well-behaved as the Pool 2 statistics, our results, in general, appear to provide good news for testing programs that wish to establish DIF screening procedures for CATs. In the case of the pretest items, the DIF statistics also appeared to be well-behaved. However, the standard errors of the Mantel-Haenszel DIF statistics tended to be larger than in the CAT, reducing the power to detect DIF.

2. Simulation procedures

Our principle in developing the simulation design was to aim for some reasonable compromise between an approach that was realistic (in that it mimicked the properties of an actual CAT) and one that was simple enough to yield useful, interpretable results. In designing the simulation, we consulted with staff from ETS testing programs to ensure that our decisions were likely to produce data that were substantially consistent with actual ETS test 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 and pretest items. These components are described in the following sections.

2.1 Administration conditions

Eighteen data sets were created, each corresponding to a CAT administration. The administrations were defined by the properties of the item pool, the ability distributions of the reference and focal groups, and the group sample sizes. These factors are described below. The number of levels of the three factors was 3, 3, and 2, respectively, resulting in 18 distinct data sets, the properties of which are summarized in Table 1.

Insert Table 1 about here.

Item pool: Three item pools were included. Pool 1 had no DIF; its purpose 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 item difficulty, and Pool 3 had DIF that was positively correlated with item difficulty. Research has found that, for some pairs of ethnic groups, DIF tends to be positively correlated with item difficulty, whereas for male-female analyses, this tends not to be true (e.g., Kulick & Hu, 1989). Pools 2 and 3 were created to allow investigation of the effect of this correlation. The item difficulty, discrimination, and guessing parameters were the same across all three

pools of items; only the DIF properties varied. Details on the pattern of DIF are given in section 2.2.

Focal group ability distribution: The three possible focal group distributions were 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 chosen 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 and focal groups, respectively. Like the focal group distributions, these sample size conditions were chosen to be similar to those that occur in ETS analyses.

2.2 CAT simulation

In simulating the CAT data, item responses were generated based on the true item parameters, using the 3PL item response function,

$$P_j(\theta) = c_j + (1 - c_j)\,(1 + \exp(-1.7a_j(\theta - b_{jG})))^{-1}, \quad (1)$$

where $P_j(\theta)$ is the probability of answering item j correctly for examinees with ability θ, a_j and c_j are the discrimination and guessing parameters, respectively, b_{jG} is the difficulty in

group G (G = reference or focal), and the factor of 1.7 is included to make the logistic scale into an approximate probit scale (Lord & Novick, 1968, p. 400). The focal group difficulty, b_{jF}, was obtained by adding the item d_j value, which could be positive or negative, to the reference group difficulty, b_{jR}. A response was generated as correct if a random number drawn from a uniform distribution between 0 and 1 was less than the value of the item response function computed at the true ability. Otherwise the response was incorrect.

The CAT simulation was designed as a simplified version of actual CATs being developed at ETS. The CAT algorithm selects as the next item to be administered the most informative item at the maximum likelihood estimate of ability computed from the items already administered.[1] (Estimates of item information and examinee ability were computed using estimated item parameters, described in section 2.3.6). Most actual CATs under development at ETS select items on the basis of both information and other characteristics, such as item format and content.

We based our study on a fixed-length CAT of 25 items. This is similar to the number of items in a single section of the SAT and GRE CATs. To determine the number of items in the CAT pool, we conducted trial CAT simulations to allow us to investigate patterns of item use for pools with various item properties. We considered pools of 75 and 100 items, and concluded that the 75-item pool was superior in that a higher percentage of the items

[1]The CAT algorithm was implemented in a revised version of a program written by Martha Stocking based on the approach of Lord (1976). The item information function is defined as $P'_j(\theta)^2 / P_j(\theta)Q_j(\theta)$, where $P_j(\theta)$ is the item response function (in this case, the 3PL function defined in equation 1), $P'_j(\theta)$ is the first derivative of $P_j(\theta)$ with respect to θ, and $Q_j(\theta) = 1 - P_j(\theta)$ (see Lord, 1980).

were actually used. This ratio of items in the pool (75) to items administered per examinee (25) is smaller than in many real applications. However, using a larger pool would have meant a reduction in the percent of pool items that were administered.

Selection of the most informative item at the examinee's estimated ability was achieved using an item information table, shown in Table A-1 in Appendix A, that contains columns for equally spaced abilities from -2 to 2 at intervals of .2. Each column lists the item numbers sorted in descending order by the item information at that ability level. The table contains 25 rows, since each CAT consisted of only 25 out of the 75 pool items. (To allow additional analysis, examinee responses were also generated for all of the pool items not administered in the CAT.)

In a process similar to that used in actual CATs, the first item administered was randomly selected from the first four items in the column at ability zero. The second item was randomly selected from the first three items at either an ability of -2 or +2, depending on whether the first item was answered incorrectly or correctly, respectively. Examinees with all-incorrect or all-correct patterns after responding to item 2 continued to receive the most informative item (among those not yet administered) from the -2 or the +2 column, respectively. Once an examinee had both a right and a wrong answer, ability was reestimated by maximum likelihood following each item response. Each subsequent item was selected from the column of the information table which was closest to the examinee's estimated ability, calculated from responses to all items answered up to that point. The most informative item that had not already been given to that examinee was administered.

Examinees that answered all CAT items incorrectly were assigned an ability estimate of -10. Examinees that answered all items correctly were given an ability estimate of 10.

Item usage for all conditions is given in Table A-2 in Appendix A. The body of the table gives the number of examinees, out of a total of 60,000 for each population group, who were administered each item in the pool. Note that four of the 75 CAT items were never administered. This occurred because, at every ability level, there were at least 25 items that were more informative than these items. This phenomenon occurs in real CATS as well. To show how the usage of items varies across ability level, two illustrative tables were produced. Table A-3 shows item usage for various ability intervals in the reference group. Table A-4 gives the corresponding information for the focal N(-1,1) group for the Pool 3 items.

2.3 Specification of item parameters

Within each of the 18 "administrations," the factors that were varied were the item discrimination (a) and reference group difficulty (b) parameters[2] and the item d parameters, representing the degree to which the item difficulties differed. Decisions needed to be made about the distributions of item parameters (assuming a 3PL item response function) and of the DIF parameters d. We chose to use multivariate normal distributions to model the joint distribution of the DIF and item parameters, with a natural log transformation applied to the a parameter. We used three different multivariate normal distributions, each corresponding to

[2]Although we use the notation b_R to represent the reference group difficulty in some instances, we suppress the subscript for simplicity of notation in others. A b without a subscript refers to the difficulty for the R group.

an item pool, to generate the items. The parameters of these distributions are given in Tables 2 and 3. Sections 2.3.1 - 2.3.4 describe how we determined the means, standard deviations, and intercorrelations shown in the tables. The parameters for the pretest items were selected in a much simpler fashion, described in section 2.3.5. Procedures for obtaining item parameter estimates for use in analysis are described in section 2.3.6.

Insert Tables 2 and 3 about here.

2.3.1 Marginal mean and standard deviation of distribution of *d*

In this study, the DIF parameter for item *j* was defined as $d_j = b_{jR} - b_{jF}$. Therefore, a value of *d* greater than zero implied that an item was easier for the focal group than for the reference group, whereas *d* less than zero implied that the item was harder for the focal group. To decide on the distribution of *d* in Pools 2 and 3, we used both theoretical and empirical findings on the relation of *MH D-DIF* to *d*.

Donoghue, Holland and Thayer (1993) used the work of Holland and Thayer (1988) to show that, under certain Rasch model conditions, the *MH D-DIF* statistic provides an estimate of $-4ad$. The assumptions under which this finding holds are that (1) within each of the groups (reference and focal), the item response functions follow the Rasch model (obtained from equation 1 by setting $c_j = 0$ for all items *j* and $a_j \equiv a$ for all items *j*) (2) the matching variable is the number-right score based on all items, *including* the studied item, and (3) the items have the same item response functions for the reference and focal groups (i.e., $b_{jR} = b_{jF} \equiv b_j$), with the possible exception of the studied item.

Previous simulation work has shown that, when guessing is present, the appropriate multiplier is less than 4. In addition to nonzero guessing parameters, our simulation study included two values of *a*, rather than a single one. To help us select an appropriate marginal mean and standard deviation of *d* for Pools 2 and 3, we examined the regression of *MH D-DIF* on *ad* for several sets of simulated data. We found the multiplicative constants to be between 2 and 3 and the additive constants to be about zero. Using this result, we were able to determine a mean and standard deviation for *d* (0 and .3) that would produce realistic distributions of *MH D-DIF*. In Pool 1, the DIFless pool, the mean and standard deviation of *d* were, of course, zero.

2.3.2 Marginal means and standard deviations of distributions of item parameters

Properties of actual data sets were used to determine how to model the marginal means and standard deviations of the item parameters. Verbal and Mathematical sections of two forms of the SAT test were obtained from College Board Statistical Analysis for this purpose. One form, 3KSA07, had not been screened based on DIF pretest information, the other form, 3LSA02, had been. We looked at the statistics for all items and for only those items that were included in the pool. From these, the means and standard deviations of *ln a*, *b*, *c*, and the *MH D-DIF* statistic (for male-female, White-Black, and White-Asian analyses) were obtained. Also, as supplementary information, the means and standard deviations of item parameters from the initial CAT pool for the GRE quantitative section were obtained.

To simplify the simulation, we set c_j equal to .15 for all items. This value was close to the average value in the SAT and GRE data sets. The means and standard deviations of *ln a* (-.15, .30) and *b* (0, .15) were also chosen to be similar to the average values for these data.

2.3.3 Intercorrelations among item and DIF parameters.

The SAT data sets described in the previous section were also used in modeling the intercorrelations among the parameters. For purposes of determining the correlation of DIF with the other parameters in Pools 2 and 3, we used the *MH D-DIF* statistic as a proxy for *d*. To aid in determining reasonable intercorrelations of *a*, *b*, and *d*, the partial correlations of the estimates of *ln a*, *b*, and *MH D-DIF*, with the estimated *c* partialed out, were examined, in addition to the zero-order correlations.

The intercorrelations of the item parameters were determined as follows. Pool 1 has no DIF, so *d* is uncorrelated with *ln a* and with *b*. By design, *d* is also uncorrelated with *b* in Pool 2, which closely resembles the actual results for male-female DIF in our SAT data sets, in the data analyzed by Kulick and Hu (1989) and in other unpublished analyses of College Board data. The correlation of *b* and *d* for Pool 3 was set equal to .40, which is approximately equal to the average of the correlations of *b* and *MH D-DIF* for the White-Black and White-Asian DIF analyses in the SAT data sets. The average correlation of *ln a* and *MH D-DIF* in the SAT data sets was .04. There was considerable variation, but it did not seem to follow a meaningful pattern. Therefore, a value of zero, which approximated the mean correlation, was assigned for all three pools. The average correlation of *ln a* and *b* was

about .40; this value was assigned in all 3 pools. Table 3 shows the correlation values used for modeling the joint distribution of *ln a*, *b*, and *d* in each of t! three pools.

2.3.4 Discretized multivariate normal approach

To generate the item parameters, we assumed a multivariate normal distribution of *ln a*, *b*, and *d*, and then discretized it so that only selected values of each parameter could occur. By discretizing the distribution, we could assure that only a finite number of item types were possible, to facilitate summarization and interpretation of results. The values of the parameters that were selected for inclusion in the study were

ln a: -.3, 0 (corresponding to *a* values of .74, 1)

b: -1.95, -1.3, -.65, 0 .65, 1.3, 1.95

d: -.70, -.35, 0, .35, .70 in Pools 2 and 3; $d = 0$ for all items in Pool 1.

This implies a total of 14 possible combinations of *a* and *b*, each of which could have five possible levels of DIF in Pools 2 and 3.

The probabilities from a multivariate normal distribution with the specified parameters were used to assign probabilities to the cells of a 2 x 7 x 5 contingency table. To understand how this was done, consider the *b* parameter. As noted above, there were to be seven values of *b*, separated by .65. The probability associated with $b = x$ was defined as $P(x - .65/2 < b < x + .65/2)$, with the following modification: Probabilities associated with values outside the intervals surrounding the desired seven values of *b* were set to zero, and the remaining probabilities were renormed so that they would sum to 1. The resulting probabilities were

then multiplied by the desired number of items for the pool and then rounded to integer values.

We generated the parameters for Pool 3 first. Pool 1 was easily obtained from Pool 3 by setting all *d* parameters to 0. Pool 2 was obtained as described above but with the added restriction that the marginals for a and *b* remain the same as for the other two pools. The generated frequencies of *d* needed to be adjusted to meet this restriction. The resulting joint distribution was then checked to verify that the correlations were close to their intended values.[3] The joint frequency distributions of CAT item parameters are given in Tables A-5, A-6, and A-7 for Pools 1, 2, and 3, respectively. The *a*, *b*, and *d* parameters for all three pools of items are given in Table A-8.

2.3.5 Nonadaptive pretest item parameters

In large testing programs, test forms often include not only items that will be used in computing the examinees's overall score, but "pretest" items that are being evaluated for possible future use. Because the items have never been administered, item parameter estimates are not available. In CAT-administered exams, some testing programs are choosing to accompany adaptively administered items with a set of pretest items that are not adaptively administered. Therefore, we wanted to consider DIF analysis procedures for such items.

---

[3]Note that, in this study, an item number (1-75) defines a combination of *a*, *b*, and *c* parameters and *a*, *b*, and *c* parameter estimates. These values are associated with that item number, regardless of item pool. However, the DIF properties of the items vary across pools. The items in Pool 1 have no DIF and the amount of DIF associated with a particular item number is not, in general, the same for Pools 2 and 3.

Responses were generated to the same set of 15 pretest items for each examinee. For these items, all values of *ln a* were equal to 0 (corresponding to $a = 1$). The five levels of *d* were crossed with three levels of *b*: -1.3, 0, and 1.3. Parameters for the pretest items are given in Table A-9. The pretest items were identical for all examinees, regardless of the pool from which the CAT items were selected. The DIF analysis method applied to pretest items is discussed in section 3.2.

2.3.6 Item parameter estimation for the CAT

The CAT item parameter estimates used for computing item information and ability estimates were obtained through an analogue to a paper-and-pencil test administration. (The administration and analysis of the pretest items did not require that item parameter estimates be obtained for these items.) A sample of 2,000 examinees were "administered" all 75 items, and the LOGIST program (Wingersky, 1983; Wingersky, Patrick, & Lord, 1988) was used to estimate the *a*, *b*, and *c* parameters for each item. Because 2,000 is a typical sample size for such calibrations, this approach allowed us to incorporate a realistic amount of estimation error. The estimated *a*, *b*, and *c* parameters, which were the same for all three pools, are given in Table A-10, along with the true parameters.

We included only members of the reference population in our calibration sample. Initially, we considered using a sample consisting of both reference and focal group members for item calibration or using a weighted combination of true reference and true focal group parameters, possibly with an error term added. However, because we wished to compare our

results across simulation conditions, it was desirable to use the same set of parameter estimates in all cases. In fact, in actual CATs, a single set of parameter estimates is used, regardless of the demographic composition of the test-takers in a particular administration. 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; therefore, including only reference group members appeared to be the best procedure. In our simulation, estimation of both item information functions and examinee ability is based on an incorrect (DIFless) model for the focal group. This closely approximates the situation that arises in actual testing applications when the true item response functions 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.

3. DIF analyses

Originally, our investigation was to focus on three general DIF approaches: (1) the MH and standardization DIF methods, using expected true score on the CAT as a matching variable, (2) a variation on (1) for nonadaptive pretest items, in which the matching variable is the sum of the expected true score on the CAT and the score on the studied item, and (3) comparison of item percents correct for late-occurring items. In addition, we planned a comparison between DIF results obtained from CATs to results obtained by administering all pool items and matching either on expected true score based on all item responses or on number-right score.

Preliminary simulations allowed us to eliminate from further consideration the method based on comparison of item percents correct for late-occurring items. The reasons for eliminating this method are described in the next section, followed by a description of the MH and standardization methods.

3.1 Comparison of item percents correct for late-occurring items

In this proposed DIF analysis method, an examinee's data for an item were to be included in the analysis only if the examinee received the item in the latter part of the CAT. Then, the simple differences in item percents correct for the reference and focal groups were to be examined. This approach was based on the expectation that examinees who received an item toward the end of their CATs would be quite well matched in ability, so that DIF statistics and simple differences in percents correct would yield similar conclusions (Holland & Zwick, 1991). However, results from simulation data indicated that this matching strategy did not work as expected. Two types of simulation data were generated. In one simulation, item parameters for a 75-item CAT pool were constructed according to the procedures described in section 2.3. Five thousand examinees were selected from a standard normal ability distribution and were administered a 25-item CAT. The mean and variability of true ability for examinees who took items late in the test were then examined. Specifically, we compared the mean and standard deviation of true ability for (1) all examinees taking the item, (2) examinees taking the item in positions 16-25, and (3) examinees taking the item in position 25. In only 32 of 75 items was the variance of ability smaller for examinees who

took the item in positions 16-25 than for all examinees taking the item. Similarly, restricting attention to only those who took the item last did not assure a decrease in variability and, of course, led to dramatic sample size reductions.

To determine whether this undesirable result was a result of artificial properties of our particular simulation, the same type of analysis was conducted using preliminary item parameter estimates for 90 items from the actual CAT pool for the GRE quantitative section. Again, results were obtained for a sample of 5,000 from a N(0,1) population. In this simulation, it was found that restricting attention to late usage (positions 16-20 for a 20-item CAT) led to variance reductions in only 40 of 90 items. Examination of the information tables for both these simulations showed that items often appeared toward the bottom portion of the table for several widely separated ability levels. The situation is likely to be exacerbated in the case of actual CATs, in which constraints on item type and content (e.g., not too many items on a particular topic) will mean that item information plays a less important role in selecting items. Based on our early simulation findings, we excluded this method from the remainder of our study.

3.2 The Mantel-Haenszel and standardization DIF procedures [4]

In both the MH and standardization methods of DIF analysis, examinees are first grouped on the basis of a matching variable that is intended to be a measure of ability in the

[4]The description of the MH D-DIF and STD P-DIF statistics is adapted from Donoghue, Holland, & Thayer (1993).

area of interest. In most DIF applications, the matching variable is a total test score, based either on the test in which the studied item is embedded or, if the studied item is being pretested, on a separate test in the same subject area. 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. This 3-way classification forms the basis of both the MH and standardization procedures. One 2 x 2 layer of this 2 x 2 x K array is represented below.

	Performance on the Studied Item		
Group	Correct = 1	Incorrect = 0	Total
Reference	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

In this notation, there are T_k examinees with the same value of the matching variable. Of these, n_{Rk} are in the reference group and n_{Fk} are in the focal group. Of the n_{Rk} reference group members, 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. The MH measure of DIF is defined as

$$MH\ D\text{-}DIF = -2.35\ ln(\alpha_{MH}) \qquad (2)$$

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

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k/T_k}{\sum_k B_k C_k/T_k} . \tag{3}$$

The transformation of $\hat{\alpha}_{MH}$ in (2) places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The effect of the minus sign in (2) 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. An estimated standard error for *MH D-DIF* is given in Holland and Thayer (1988), based on work reported in Robins, Breslow and Greenland (1986) and Phillips and Holland (1987). It is

$$SE(MH\ D\text{-}DIF) = 2.35\sqrt{Var(ln(\hat{\alpha}_{MH}))} \tag{4}$$

where $Var(ln(\hat{\alpha}_{MH}))$ is estimated by

$$\frac{\sum_k U_k V_k/T_k^2}{2(\sum_k A_k D_k/T_k)^2} , \tag{5}$$

where

$$\begin{aligned} U_k &= (A_k D_k) + \hat{\alpha}_{MH}(B_k C_k) \\ V_k &= (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k). \end{aligned} \tag{6}$$

The Mantel-Haenszel chi-square test of the null hypothesis of no difference between the performance of the focal group and of comparable members of the reference group on the studied item was not examined in our study.

The standardization DIF measure, developed by Dorans and Kulick (1986), is

$$STD\ P\text{-}DIF = \hat{p}_F - \tilde{p}_R \tag{7}$$

where $\hat{p}_F$ is the proportion in the focal group who get the studied item correct, and $\tilde{p}_R$ is an adjusted proportion correct on the item for the reference group, defined as

$$\tilde{p}_R = \sum_k \left(\frac{A_k}{n_{Rk}}\right) \frac{n_{Fk}}{n_F} \tag{8}$$

where $n_F = \sum_k n_{Fk}$ is the total number of examinees in the focal group. One interpretation of $\tilde{p}_R$ is that it is the proportion of reference group examinees who would have got the studied item right had the distribution of the matching variable in the reference group been the same as it is for the focal group.[5]

The estimated standard error for *STD P-DIF* is given by the formula

$$SE(STD\ P\text{-}DIF) = \sqrt{\sigma_F^2 + \hat{\sigma}_R^2} \tag{9}$$

---

[5]When n_{Rk} is equal to zero, both $\tilde{p}_R$ and σ_R^2 are undefined. When this occurs, the standard ETS DIF software implements an imputation procedure proposed by Holland (McHale, Dorans, Holland, & Petersen, 1988). Analogous procedures, modified to take into account the special nature of the CAT-based analyses, were used in our work.

where

$$\sigma_F^2 = \frac{1}{n_F} \hat{p}_F (1 - \hat{p}_F) \tag{10}$$

and

$$\sigma_R^2 = \frac{1}{n_F^2} \sum_k \frac{n_{Fk}^2 A_k B_k}{n_{Rk}^3} . \tag{11}$$

Our matching variable for the DIF analysis of the CAT-administered items was obtained by (1) getting the examinee's MLE of ability, 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 item response functions. That is, the matching variable was

$$\textit{Expected true score based on CAT} = \sum_{j=1}^{75} \hat{P}_j \left(\hat{\theta}_{CAT}\right), \tag{12}$$

where $\hat{P}_j(\cdot)$ is an estimate of the function defined in equation 1 and $\hat{\theta}_{CAT}$ is the MLE of ability based on the CAT items. Examinees whose expected true scores fell in the same one-unit intervals were considered to be matched. For the pretest items, which were administered nonadaptively, the matching variable was the sum of the expected true score on the CAT, computed according to equation 12, and the score (0 or 1) on the studied pretest item.

4. Organization of the data for DIF analysis

4.1 Examinee records

The record that was constructed for each examinee contained the following information: population indicator (either reference or one of 3 types of focal), true ability, pool indicator (1, 2, or 3), string of 75 responses to all pool items, item numbers of CAT items administered (in order), string of 15 pretest item responses, estimated ability for the 75-item nonadaptive test and for the 25-item CAT, and expected true scores corresponding to each of the two ability estimates.

Generating responses to all 75 pool items had two purposes: (1) These responses could be used for the "nonadaptive control" part of the study, which attempted to distinguish the effects of using CAT data from the effects of using expected true score as a matching variable and (2) the responses could be used to construct additional CATS for the examinees if desired by using the CAT algorithm to generate a CAT sequence and plugging in the existing item responses. Although we did not make use of (2), constructing the record in this way makes it possible for us to generate data less expensively in future research.

4.2 Definition of sample size conditions

In our CAT setting, 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 huge range. For example, in Table A-2, the range of item sample sizes is from 0 to 51,133 (out of a total of 60,000) for the focal group in Conditions 3 and 4. Because our goal was to investigate the behavior of selected DIF statistics under a fixed sample size, simply analyzing the available data for each item was clearly undesirable. We therefore considered several other approaches. Initially, we attempted to generate enough data to meet the target item sample sizes for all conditions. This implied that 900 reference group members were needed, along with 500 members of each of the three focal groups for each of the three pools (see Table 1). To achieve this goal for most items required generating 60,000 cases for the reference group and for each of the nine focal distribution by item pool combinations. To assess variability, we planned to conduct two replications per condition.

After examining the DIF results from this approach, we concluded that the standard errors of the DIF statistics were large enough to make it difficult to characterize the behavior of the statistics for different item types and different conditions. Even averaging across two replications did not appear adequate. Because of the cost of data generation, we did not wish to simulate additional data. We considered several resampling approaches, which would have allowed us to obtain multiple estimates of each statistic, but none seemed ideal for our purpose. The approach we ultimately decided to use, proposed by Charles Lewis, was as follows: For each item, we used all the available CAT data (out of a maximum of 60,000 responses per group) to form the 2 (item responses) x 2 (groups) x K (levels of the matching variable) contingency table needed for DIF analysis (see section 3.2). We then converted the table frequencies to proportions of the total number of observations. Using these proportions

as estimates of the population probabilities associated with the 4 x K cells for the particular configuration of conditions in question, we obtained expected tables for our target sample sizes by multiplying the probability estimates for focal group cells by the desired focal group sample size and then doing the same for the reference group cells. Next, we computed DIF statistics and standard errors, based on the expected tables, for all 18 conditions in Table 1. (Note that the estimate of the *STD P-DIF* statistic obtained using the expected table approach is the same as the value obtained using all available data, regardless of the target sample sizes.)

As a simple example of the expected table (ET) approach, consider the following hypothetical data for a single item, assuming that there are only two levels of the matching variable. The first step would be to use all the data available for the item to construct a 2 x 2 x 2 frequency table (because $K = 2$ here). Then the frequencies for the reference group would be divided by the total number of reference group examinees and the frequencies for the focal group would be divided by the total number of focal group examinees, producing the following 2 x 2 x 2 table of probabilities:

Low on Matching Variable

	Right	Wrong	Total
Reference	.2	.1	.3
Focal	.2	.2	.4

High on Matching Variable

	Right	Wrong	Total
Reference	.5	.2	.7
Focal	.4	.2	.6

Now assume that we wanted target tables for the $n_R = 900$, $n_F = 100$ condition. The reference group probabilities would be multiplied by 900 and the focal group probabilities would be multiplied by 100, producing the following table, which would then be used for DIF analysis.

	Low on Matching Variable		
	Right	Wrong	Total
Reference	180	90	270
Focal	20	20	40

	High on Matching Variable		
	Right	Wrong	Total
Reference	450	180	630
Focal	40	20	60

For the *MH D-DIF* statistic, the formula for the standard error of the ET estimate, $SE_{ET}(MH\ D\text{-}DIF)$, is given in Appendix B. For the sample size conditions we investigated, its value is very similar to the value of *SE*(*MH D-DIF*) (equations 4-6) obtained using all the available data. For the *STD P-DIF*, the standard error of the ET estimate, $SE_{ET}(STD\ P\text{-}DIF)$, is identical to the value of *SE*(*STD P-DIF*) (equations 9-11) obtained using all the data; therefore, no special computing formula is required. $SE_{ET}(MH\ D\text{-}DIF)$ and $SE_{ET}(STD\ P\text{-}DIF)$ are typically much smaller than the ordinary standard errors that would be obtained for the target sample sizes in question.[6] Therefore, even though it produces only a single estimate,

[6] $SE_{ET}(MH\ D\text{-}DIF)$ and $SE_{ET}(STD\ P\text{-}DIF)$ reflect the degree of precision with which the population value is estimated using the ET approach. Because the ET estimates are typically based on thousands of cases in this study, these standard errors tend to be small. They should not be confused with the standard errors that are computed based on the expected tables generated with the ET approach, using the usual formulas (equations 4-6 and equations 9-11). This second type of standard error (which does not have an "ET" subscript) closely

the ET approach can provide a relatively precise idea of the behavior of the DIF statistics. A comparison of the ET method to an estimation procedure based on multiple replications appears in Appendix C. Our comparison was based on pretest items, for which 60,000 responses per population group were available for each item. As shown in Table C-1, the ET method was found to give results similar to those of the replication-based approach. For the items we studied, the ET estimate of *MH D-DIF* was as precise as the average over 316 replications of the *MH D-DIF* statistic based on the target sample sizes. Another advantage of the ET approach is that, once the 2 x 2 x *K* probability tables have been created, DIF results can be generated easily for any target sample size. This will be useful if we wish to consider other sample size conditions in the future.

5. Results

The results of the study are summarized in the following sections. Section 5.1 gives results for the items in the 75-item CAT pool, section 5.2 gives results for the pretest items, and section 5.3 gives some results on ability estimation for examinee groups.

5.1 Results for the CAT pool items

Results are given first for the comparison of CAT-based DIF results to nonadaptive DIF analyses. Correlations between CAT-based DIF statistics, DIF statistics based on

---

approximates the standard errors that would be obtained using actual samples of the target sizes.

nonadaptive administration, and "true DIF" are presented first, along with the means of the various DIF measures. For purposes of this analysis, true DIF was defined as the product of the item discrimination parameter (a) and the difference between the item difficulties for the reference and focal groups (d). The theoretical rationale for defining true DIF in this way is given in section 2.3.1. Following this, tables of *MH D-DIF* and *STD P-DIF* means for every combination of ad and b are given, along with a discussion of the standard errors of the DIF statistics. Finally, an estimate is given of the proportion of times each type of item would be declared an extreme DIF ("C") item using the ETS method of classifying items into DIF categories.

5.1.1 Comparison of CAT-based and nonadaptive DIF analyses

For selected simulation conditions, we compared MH and standardization results from the CAT analyses, described above, to results of two nonadaptive DIF analyses. The first was a procedure ($\hat{\theta}$-75) in which all 75 pool items were "administered" and examinees were matched on expected true score calculated using the MLE of ability based on all 75 responses (the "nonadaptive control"). That is, instead of the matching variable in equation 12, the matching variable was

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

where $\hat{\theta}_{75}$ is the MLE of ability based on all 75 items. The second approach (NR) was a conventional DIF analysis, in which all 75 pool items were administered and examinees were matched on number-right score. The results of this comparison are given in Tables 4-6.

Insert Tables 4-6 about here.

For this analysis, we chose to include only the simulation conditions that had DIF and were based on reference and focal sample 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 types of DIF statistics and the true DIF for the item. Each correlation matrix was based on the 71 items that were administered in the CATs (see section 2.3).

The CAT-based *MH D-DIF* and *STD P-DIF* statistics used in this analysis were computed using the ET method, while the two other statistics were computed based on actual samples of 500 from the reference and focal 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, we computed correlations that were corrected for unreliability, using the following formula:

$$r_{XY}^{C} = \frac{r_{XY}}{\sqrt{r_{XX} \cdot r_{YY}}} \tag{14}$$

where r_{XY}^{C} is the corrected correlation between X and Y, r_{XY} is the ordinary Pearson correlation between X and Y, and r_{XX} and r_{YY} are the reliabilities of X and Y. For a particular type of DIF statistic (*MH D-DIF* or *STD P-DIF*), reliability was estimated as

$$Reliability = 1 - \frac{\sum_j SE_j^2\,(DIF\ statistic)/J}{Variance\ across\ J\ items\ of\ DIF\ statistic}, \qquad (15)$$

where J is the number of items. The numerator on the right-hand side represents error variance, while the denominator represents total variance. (For the CAT DIF statistics, the $SE^2(\cdot)$ values were the squares of the $SE_{ET}(MH\ D\text{-}DIF)$ or $SE_{ET}(STD\ P\text{-}DIF)$ values as appropriate; see footnote 6. The reliability of *ad* is, of course, unity, since it is not a statistic.) These corrected correlations provide a more equitable way of comparing the CAT, $\hat{\theta}$-75, and NR analyses.[7]

Both uncorrected and corrected intercorrelations of the values of the *MH D-DIF* statistic for the three types of matching variables and the values of the true DIF are given in Table 4 for each of the six conditions. The median across conditions is also given. The corresponding information for *STD P-DIF* is given in Table 5. Both Tables 4 and 5 show that the CAT, $\hat{\theta}$-75, and NR analyses produced results that were highly correlated with each other and with the true DIF values. In particular, the two analyses based on all 75 item responses produced virtually identical results (with corrected correlations exceeding unity). The median corrected correlation with true DIF was about the same for the CAT, $\hat{\theta}$-75, and NR analyses, which is somewhat surprising since the CAT DIF approach matches examinees on the basis of only 25 item responses. In general, correlations tended to be slightly higher for the *MH D-DIF* statistics than for *STD P-DIF*. The near-unity correlations of the CAT DIF statistics with true DIF was a welcome finding.

[7]If reliability is underestimated, the corrected correlation in equation 14 can exceed unity.

For some simulation conditions, we have direct evidence that use of the ET method gives similar correlation results on the performance of the CAT-based DIF procedure as does an analysis based on actual samples of the target sizes. For Condition 6, we compared MH results based on actual samples of 500 per group to the ET results. Based on the samples of 500, the uncorrected correlations of the CAT *MH D-DIF* statistics with *MH D-DIF* values from the $\hat{\theta}$-75 and NR 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 *MH D-DIF* statistics with true DIF was .92, compared to .95 for the ET method.

High correlations alone, however, do not ensure the accuracy of the DIF methods. To determine whether the obtained statistics had the desired means, we computed, for each analysis strategy in each simulation condition, the mean *MH D-DIF* and *STD P-DIF* values across the 71 items that were given in the CAT, along with the standard deviation across items. The results are given in Table 6, along with the medians over the six simulation conditions. (In the case of *STD P-DIF*, means and standard deviations have been multiplied by 10.) The mean across 71 items of the true DIF values is -.004 in Conditions 4, 10 and 16 (Pool 2) and -.001 in conditions 6, 12, and 18 (Pool 3), with a standard deviation of .293 in both pools.

In *MH D-DIF* analysis in which all examinees take all items and the matching variable is number-right score, the average *MH D-DIF* is constrained to be approximately zero across items, producing a negative covariance among the DIF statistics within a test. If it were not for rounding error and for the adjustment procedure described in Footnote 2, the *STD P-DIF* statistics would sum to zero under these conditions as well. This constraint on the *MH D-*

DIF and *STD P-DIF* is not present in the CAT and $\hat{\theta}$-75 analyses. In these other 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 6 for two reasons: First, because of the constraint on the mean of the NR-based statistics, it is not clear which across-item NR mean is 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 accompanying standard deviations) are therefore included in Table 6. Second, the non-zero covariances for the NR-based statistics and possibly for the other analyses makes it difficult to estimate the standard errors of the means in Table 6. If the *MH D-DIF* statistics were independent across items, the standard errors of the average *MH D-DIF* statistics in Table 6 would be roughly .009 for the CAT analysis and .049 for the two nonadaptive analyses (obtained by dividing the average item-level standard error by the square root of the number of items). Judged in this light, the means for the nonadaptive procedures were quite close to zero, but the means for the CAT procedure were slightly inflated. All six means for the CAT-based procedure were greater than zero and the means were larger for the Pool 3 conditions than for the Pool 2 conditions. However, these values for the standard error of the mean are only approximate. Because of the negative covariances among NR *MH D-DIF* statistics within a test, the value of .049 is definitely an overestimate of the standard error of the mean for the NR analyses. Presumably, this overestimation holds for the $\hat{\theta}$-75 approach, which produced results nearly identical to the NR analyses. For the CAT DIF statistics, the value computed under independence may either under- or overestimate the standard error. In any case, the practical implications of an inflation of .01 to .05 in the *MH D-DIF* statistic are

small in that a difference this size is unlikely to have much effect on decisions about the item. (Of course, it would be possible to rescale the statistics so that they would be centered on zero for a particular collection of items.) Under independence, the standard errors of the means for *STD P-DIF* x 10 in Table 6 would be about .008 for the CAT analysis and .03 for the two nonadaptive approaches (obtained, once again, by dividing the average standard error by the square root of the number of items). Again, there appears to be a slight inflation of the statistics in the CAT analysis. There were also relatively large departures from zero for the two nonadaptive methods in Condition 4.

In addition to comparing the values of *MH D-DIF* and *STD P-DIF* for the three matching variables, we also examined their standard errors. For the $\hat{\theta}$-75 and NR analyses, the average values of *SE(MH D-DIF)* within each condition were about .40, whereas the CAT-based *MH D-DIF* statistics tended to have standard errors of about .35. One hypothesis for this discrepancy is that the smaller standard errors for the CAT DIF analysis are related, at least in part, to the use of the ET estimation method. Table C-1 shows that the ET-based estimates of *SE(MH D-DIF)* tended to be smaller than the average *SE(MH D-DIF)* across 66 replications by about .03. Another hypothesis is that CAT-based methods of DIF analysis tend to produce lower standard errors for reasons unrelated to ET estimation, such as the restriction of the analysis to examinees in a smaller ability range (see the related discussion in section 5.2.2). The interpretation of the standard error results for the three DIF analyses is complicated, however, by the fact that the pattern of standard errors for *MH D-DIF* is not paralleled by the results for *SE(STD P-DIF)*, where average standard errors (multiplied by 10) ranged from .25 to .32. Here, the average *SE(STD P-DIF)* within a condition did not vary

much across the $\hat{\theta}$-75, NR and CAT DIF statistics. Although the differences were small, however, the values of *SE(STD P-DIF)* were *larger* for the CAT approach than for the $\hat{\theta}$-75 and NR approaches for all six simulation conditions. These standard error findings, along with those described in section 5.2.2, require further investigation.

5.1.2 *MH D-DIF* and *STD P-DIF* statistics by item type

In addition to comparing the CAT approach to nonadaptive DIF analysis methods, we examined the average CAT DIF statistics for various types of items and simulation factors. To determine the best way to summarize the results, we conducted a series of analyses of variance (ANOVAs) in which the observations were the DIF statistics and the independent variables were sample size condition, focal group distribution, item difficulty level (B), item discrimination (A), item DIF level (D), 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). We began with the *MH D-DIF* statistics, which we 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 B effect was significant at an α of .01;[8] it explained less than 3% of the variance in the *MH D-DIF* statistics. In Pools 2 and 3, D explained about 85% of the variance. Most analyses of Pools 2 and 3 showed very

[8]As in all exploratory analyses, significance testing can be viewed here only as a rough tool for ranking the size of effects.

small, but statistically significant effects of B, and of the B x D and A x D interactions. Somewhat surprisingly, focal group distribution had, essentially, no effect, nor did it interact with other variables. Sample size too, had no effect. (Since 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 standard errors of the DIF statistics, discussed below.) Item position and pool never yielded statistically significant main effects, though these factors sometimes showed tiny interactions with other variables. We conducted similar analyses for the *STD P-DIF* statistics and obtained nearly identical results. Based on the ANOVA findings, we displayed *MH D-DIF* and *STD P-DIF* averages for every combination of *ad* and *b*.

5.1.2.1 *MH D-DIF* results

The average *MH D-DIF* statistics are given in Tables 7-9 for Pools 1, 2, and 3, respectively. Results are given for the n_R = 500, n_F = 500 sample size condition only. As noted, results were nearly identical for the two sample size conditions. The average standard error of the estimate, SE_{ET}*(MH D-DIF)*, is given as well. As described earlier, computation of the standard error of the mean DIF statistics is not straightforward. The average value of SE_{ET}*(MH D-DIF)*, given in parentheses in Tables 7-9, is the maximum value that the standard error of the mean *MH D-DIF* could take (i.e., the value that would occur if all items had intercorrelations of one) and therefore yields an overestimate of the standard error of the

mean. The third entry in each cell in Table 7-9 is the number of item results contributing to the average. Since results are averaged over the three focal group distributions, a single item within a pool generates three entries in the table in which it occurs. The total number of entries in each of Tables 7-9 is 3 (focal group distributions) x 71 (items administered in the CAT) = 213.

As shown in Table 7, the *MH D-DIF* statistics were well-behaved in the null case; they were equal to zero at the tabled level of accuracy. The bottom margins of Tables 8 and 9 show that the average value of *MH 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 (compared to the theoretical value of 4*ad* that holds in the Rasch case described in section 2.3.1). In Pool 2, Table 8 shows that, for a fixed value of *ad*, the average *MH D-DIF* usually decreased in absolute value as *b* increased. For example, for *ad* = -.35, the average *MH D-DIF* was -1.3 for *b* = -1.95, -1.2 for *b* = 0, and -0.7 for *b* = 1.95. This phenomenon, noted by Donoghue, Holland, and Thayer (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 harder the groups are to differentiate. Superimposed on this phenomenon, Pool 3 (Table 9) included a correlation between the difficulty and DIF parameters. Easier items in Pool 3 are more likely to have negative DIF than harder items. The relation between *MH D-DIF* and *b* for fixed *ad* was not as evident in Pool 3 as it was in Pool 2. Also, the average *MH D-DIF* for the DIFless items were not as close to zero as they were in Pools 1 and 2. For *d* = 0, the average *MH D-DIF* decreased from 0.3 to -0.5 as *b* increased from -1.95 to 1.95. One item that showed surprising behavior

in Pool 3 was item 75, which appears in the bottom row of the body of Table 9 at the far right. The average *MH D-DIF* departs considerably from 3*ad*= 2.10. At first, we hypothesized that this was because item 75 was one of the items that were administered randomly to some examinees at the beginning of the CAT (see section 2.2). However, we found that item 75 had an unusually small *MH D-DIF* value even when administered nonadaptively. The most likely explanation is that the small DIF value is related to the extreme difficulty of the item.

Insert Tables 7-9 about here.

The values of *SE(MH D-DIF)* varied little across pools, DIF levels, item difficulty, or item discrimination. The primary determinant of *SE(MH D-DIF)* was sample size. For the n_R = 500, n_F = 500 condition, *SE(MH D-DIF)* ranged from about 0.3 to 0.4; for the n_R = 900, n_F = 100 condition, the range was from about 0.5 to 0.7.

5.1.2.2 *STD P-DIF* results

STD P-DIF results are given in Tables 10-12 for Pools 1, 2, and 3, respectively. The *STD P-DIF* statistics, as well as the values of SE_{ET}*(STD P-DIF)*, have been multiplied by 10. Results are given for the n_R = 500, n_F = 500 sample size condition only. As noted earlier, ET estimates of the *STD P-DIF* statistics do not depend on the target sample sizes; therefore, the results for the n_R = 900, n_F = 100 sample size condition were identical. The average value of values of SE_{ET}*(STD P-DIF)* (x 10) is given as the second entry in each cell of the tables. As

noted, this value yields an overestimate of the standard error of the mean. The third entry in each cell in Tables 10-12 is the number of item results contributing to the average. The third entries are the same as those in the *MH D-DIF* results (Tables 7-9).

Table 10 shows that, in Pool 1, the *STD P-DIF* statistics were close to zero, as desired. The bottom margins of Tables 11 and 12 show that the average values of *STD P-DIF* x 10 were roughly 2.7 times the value of *ad* in Pool 2, and 2.5 times the value of *ad* in Pool 3. Table 11 shows that, unlike *MH D-DIF*, *STD P-DIF* did not tend to decrease in absolute value as *b* increased for a fixed value of *ad*. An aspect of the results that did mirror the *MH D-DIF* results was that the average *STD P-DIF* for the DIFless items in Pool 3 (Table 12) were not as close to zero as they were in Pools 1 and 2. For $d = 0$, the average value of *STD P-DIF* x 10 was 0.21 at $b = -1.95$ and 0.22 at $b = -1.30$. It then decreased as *b* increased, reaching -0.46 for $b = 1.95$. Also, as in the *MH D-DIF* results, item 75 in Pool 3, which appears in the bottom row of the body of Table 12 at the far right, had a smaller DIF statistic than expected.

Insert Tables 10-12 about here.

The values of *SE(STD P-DIF)* varied little across pools, DIF levels, item difficulty, or item discrimination. As in the case of *SE(MH D-DIF)*, the primary determinant of *SE(STD P-DIF)* was sample size. For the $n_R = 500$, $n_F = 500$ condition, *SE(STD P-DIF)* x 10 was always about 0.3; for the $n_R = 900$, $n_F = 100$ condition, it was about 0.5.

5.1.3 Estimated percent of "C" results for item types

ETS has a system for categorizing the severity of DIF based on MH results. According to this classification scheme, a "C" categorization, which represents extreme DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = .05$). A "B" categorization, which indicates moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = .05$) and that the absolute value of *MH 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.

Because most of the ET estimates of *MH D-DIF* and *SE(MH D-DIF)* statistics are based on at least 10,000 observations, it is reasonable to assume that they provide precise estimates of the population mean and standard deviation of the *MH D-DIF* statistic for the relevant configuration of item properties and simulation conditions. This is supported by the analysis described in Appendix C. If it is assumed that the *MH D-DIF* statistics for this configuration are normally distributed with this mean and standard deviation, percentiles of the *MH D-DIF* distribution can be obtained. These percentiles can then be used to estimate the percent of times such an item will be classified as an A, B, or C item.[9] This is an alternative way of providing information about the sampling variation of the *MH D-DIF* statistic.

[9]This approach was suggested by Charles Lewis.

Based on the ETS DIF rules, we developed an algorithm for estimating these percents, to be applied separately to each item in each condition (see Appendix D). The algorithm was tested and found to work well with data for 15 items from our simulation, using the ET estimates of *MH D-DIF* and *SE(MH D-DIF)* to approximate the mean and standard deviation of the *MH D-DIF* distribution. Details and results are given in Table C-1 in Appendix C. The algorithm was also tested on data from the simulation study of Donoghue, Holland, and Thayer (1993) consisting of 100 replications of the *MH D-DIF*, *SE(MH D-DIF)* and *MH chi-square* statistics for six different items. For each item, the estimated percents of A, B, and C results based on the method of Appendix D (using the average over 100 replications of *MH D-DIF* and *SE(MH D-DIF)* to estimate the mean and standard deviation of the *MH D-DIF* distribution) matched very closely the actual percents of A, B, and C results in the 100 replications.

Tables 13-15 give the average expected percent of C results for each combination of *ad* and *b*. The first entry in each cell is the average expected percent of C results for the $n_R = 900$, $n_F = 100$ condition, the second entry is the average expected percent for the $n_R = 500$, $n_F = 500$ condition, and the third entry is the number of item results contributing to the average.

Insert Tables 13-15 about here.

Table 13 shows that the percents of C results were close to zero in the null case, as desired. Even in the worst case (b = -1.95, n_R = 500, n_F = 500), the average expected percent of C results was only 0.2. As noted earlier, an item must have an *MH D-DIF* with a

magnitude exceeding 1.5 in order to be a C item. Because *MH D-DIF* was found to be approximately equal to 3*ad* in the conditions investigated in our study, items with $ad = \pm.70$ and $ad = \pm.52$ can be regarded as nominal C items. The bottom margin of Table 14 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 conditions. Table 15 shows smaller detection rates for the nominal C items in Pool 3. As noted, item 75, which has a difficulty of 1.95, had a smaller *MH D-DIF* value than expected; therefore, its average expected percent of C results was also smaller. The three items with $ad = \pm.52$ also had considerably smaller detection rates than the $ad = \pm.52$ items in Pool 2.

5.2 Results for the pretest items

For several reasons, results for the pretest items must be interpreted differently from the results for the CAT items. First, the pretest items were administered nonadaptively. Second, the pretest items were identical for all examinees, regardless of which CAT pool was administered. Therefore, the identity of the CAT pool is relevant only because the matching variable for the pretest items was a function of the expected true score on the CAT. Third, the properties of the pretest items follow a balanced design. Specifically, three levels of *b* were crossed with five levels of *d*, and *a* was equal to 1 for all items.

5.2.1 *MH D-DIF* and *STD P-DIF* for pretest items

To determine how to summarize the results of the pretest items, an ANOVA was conducted using the *MH D-DIF* statistics as the observations. (Because analyses on the CAT pool items showed that ANOVAs of *MH D-DIF* and *STD P-DIF* led to nearly identical results, no ANOVA was conducted for *STD P-DIF* for the pretest items.) The factors included were focal group distribution (F), item pool (P), item difficulty (B), and item DIF level (D). All two-factor and three-factor interactions were also assessed. Results were somewhat different from those obtained for the CAT items. All effects were statistically significant at $\alpha = .01$ except for the P x D, F x P x D, F x B x D, and P x B x D interactions. However, the only effects that explained more than 1% of the variance were D (92%) and B x D (3%). Therefore, for simplicity, results were tabled in the same way as the CAT item results; that is, results were displayed for each combination of $ad = d$ and b.

Results for the pretest items are given in Tables 16-20. Although the pretest items were not subject to problems of variability in sample sizes across items, we used the ET estimation procedure for these items, as for the CAT items. Because the *MH D-DIF* and *STD P-DIF* statistics were nearly identical across pools, results for these statistics are given for Pool 1 only (Tables 16 and 17). Results are shown only for the $n_R = 500$, $n_F = 500$

condition.

Insert Tables 16-20 about here.

In all three pools, items without DIF had DIF statistics of about zero, as desired. For items with DIF, *MH D-DIF*, but not *STD P-DIF*, tended to decrease in absolute value as *b* increased for a fixed value of *ad*. The DIF statistics for the pretest items tended to be slightly smaller than the corresponding statistics for the CAT items.

As an additional check on the DIF results for the pretest items, the correlation matrix for *MH D-DIF*, *STD P-DIF* and *ad* was obtained within each of the three pools. The three correlation matrices were nearly the same. The correlation between the two DIF statistics was .95, the correlation between *MH D-DIF* and *ad* was .96, and the correlation between *STD P-DIF* and *ad* was .94 to .95. (Because all the DIF statistics for pretest items were based on the ET method, reliabilities were close to unity. Therefore, the corrected correlations obtained using equation 14 were almost identical to the uncorrected correlations.)

5.2.2 *SE(MH D-DIF)* and *SE(STD P-DIF)* for pretest items

The most pronounced difference between the pretest and CAT item results was the size of the standard error of *MH D-DIF*. While the values of *SE(STD P-DIF)* x 10 for pretest items were, on the average, slightly smaller than those for CAT items (0.2 to 0.3 for the n_R = 500, n_F = 500 condition and 0.3 to 0.5 for the n_R = 900, n_F = 100 condition, compared to fairly consistent values of 0.3 and 0.5, respectively, for the two sample size conditions in the

CAT), the standard errors of *MH D-DIF* tended to be considerably larger for the pretest items than for the CAT items (ranging from about 0.4 to 0.6 for the $n_R = 500$, $n_F = 500$ condition and from 0.6 to 1.1 for the $n_R = 900$, $n_F = 100$ condition, compared to 0.3 to 0.4 and 0.5 to 0.7, respectively, for the two sample size conditions in the CAT).

There are several factors that may have contributed to the larger standard errors. First, they may be related to the larger group differences in ability distributions in the pretest compared to the CAT. The pretest items are administered to all examinees, whereas the CAT is administered to those within a relatively narrow range of ability. Therefore, the pretest data are distributed across more levels of the matching variable, and this greater sparseness may lead to inflation of the standard errors (see the related discussion in section 5.1.1). A second possible reason for the larger standard errors is the definition of the matching variable in the pretest analyses. The estimated standard error of *MH D-DIF* has been found to be inflated by inclusion of the studied item in the matching variable (Donoghue, Holland, & Thayer, 1993). Therefore, it is possible that the larger estimated standard errors in the pretest analyses resulted from the nonstandard method of including the studied item (i.e., adding the studied item score to an expected true score based on the remaining items). A third possible contributing factor is the relation between item difficulty and *SE(MH D-DIF)*. This phenomenon was also investigated by Donoghue, Holland, and Thayer (1993), who studied the behavior of the ratio of the average *SE(MH D-DIF)* over 100 replications to the standard deviation of *MH D-DIF* over replications. They found that this ratio was larger for items with difficulty (b) parameters of -.5 and +.5 than for those with $b = 0$. Our examination of their data revealed that the average *SE(MH D-DIF)*, like the ratio, was larger for the lower

and higher difficulty levels than for the middle difficulty level. The standard deviation of *MH D-DIF* had the opposite pattern: It was smaller for b = -.5 and b = +.5 than for b = 0. In investigating the pretest items that had unusually large standard errors, we found that these items had very large percents correct (above 85). In every simulation condition, items 4 and 5 had the largest values of *SE(MH D-DIF)*. These were the easiest of the pretest items, with b = -1.3 and d = .35 and .70, respectively. A detailed examination of pretest items for examinees in Condition 17 showed that the Spearman correlation between item percent correct and *SE(MH D-DIF)* was .88. In other conditions, such as Condition 5, the relation took a curvilinear form, which is consistent with the findings of Donoghue, Holland, and Thayer (1993). In general, whether the relation was curvilinear or monotonic, the items with the highest percents correct tended to have the highest values of *SE(MH D-DIF)*. Because the CAT items were administered to examinees with a narrower range of ability, they rarely had percents correct over 75, which may, in part, explain their smaller standard errors.

5.2.3 Expected percent of C results for pretest items

The average expected percent of C results, given in Tables 18-20 for the three pools, was, of course, affected by the larger values of *SE(MH D-DIF)* for the pretest items, as well as the slight tendency of the DIF measures themselves to be slightly smaller in the pretest than in the CAT. Results were quite similar for Pools 1 and 2, but were somewhat different for Pool 3 because of a slightly different pattern of standard errors for that pool. Given a particular value of ad and b, an item was more likely to be labeled a C item if it was a CAT

item than if it was a pretest item. Consider a Pool 3 item with $b = 1.3$ and $ad = .70$. A CAT item with these properties would be expected to have a C label 92.4% of the time in the $n_R = 500$, $n_F = 500$ condition and 54.7% of the time in the $n_R = 900$, $n_F = 100$ condition. The corresponding percents for a pretest item were 45.9% and 24.9%.

5.3 Results of examinee ability estimation

In addition to determining DIF results for groups of items, it is useful to examine the accuracy of estimation of examinee ability under various conditions. Table 21 gives, for each item pool and population group, the median and interquartile range of the residual obtained by subtracting the true ability used in data generation from the CAT-based ability estimate ($\hat{\theta}_{CAT}$). Table 22 provides the same information for the ability estimate based on responses to all 75 items ($\hat{\theta}_{75}$). Because ability estimates for examinees with infinite MLEs have been set to ±10, means and standard deviations would be misleading. Each cell of these tables is based on 1,000 examinees. The standard error of the medians are about .02 for the CAT ability residuals (Table 21) and about .01 for the 75-item ability residuals (Table 22).

Insert Tables 21-22 about here.

The most striking finding in Tables 21-22 is that all the median residuals were negative. This appears to be the result of estimation bias due to the use of estimated, rather than true item parameters. A CAT simulation based on the true item parameters showed that use of our particular set of item parameter estimates led to a downward bias in the ability

estimates for reference group examinees. In general, however, the size of the bias is not easily characterized. The bias depends on the location of the population distribution and on the presence of DIF; therefore, it is not constant across the cells of Tables 21 and 22.

Another finding was that, as expected, the median residuals were nearly always closer to zero for $\hat{\theta}_{75}$. Only in the Pool 3, focal N(+.5, 1) cell was the median residual for $\hat{\theta}_{CAT}$ slightly smaller in absolute value than the corresponding value for $\hat{\theta}_{75}$. The item-level DIF results, however, suggest that the slightly better ability estimation achieved by using all 75 item responses did not substantially improve the behavior of the DIF statistics. Certain other results are difficult to interpret. For example, estimation appears to have been better in Pool 2 than in Pool 1, which is surprising, given that Pool 1 is free of DIF.

Estimation was worst in Pool 3, particularly for the focal N(-1, 1) group, where ability was underestimated by an average of nearly one tenth of a standard deviation (of true ability) in the CAT and by about one sixth of a standard deviation on the nonadaptive test. For the CAT, this is consistent with predictions, because, in Pool 3, the lowest-ability focal group gets the easiest items, which tend to have more negative DIF. The median residual was closer to zero for the N(0, 1) population and still closer for the N(+.5, 1) group. It is interesting that in the nonadaptive administration, the pattern of median residuals for Pool 3 paralleled the pattern observed for the CAT: The N(-1, 1) focal group again had the largest median residual, followed in order by the N(0, 1) and N(+.5, 1) groups. The explanation may be that, even though all examinees receive all items in the nonadaptive administration, the most informative items are those that have difficulties close to the examinee's ability level. In the N(-1, 1) focal group, for example, the items that contribute the most to an examinee's

score are the easier items, which, in Pool 3, are more likely to have negative DIF. Other factors, such as the differential biases due to item parameter estimation, may have also contributed to the large Pool 3 residuals.

6. Summary and discussion

Our study was based on modified versions of the *MH D-DIF* and *STD P-DIF* statistics for both computer-adaptive test items and nonadaptively administered "pretest" items. We eliminated from consideration a proposed DIF method based on comparison of item percents correct for examinees who received the items late in the CAT. A preliminary simulation showed that this method did not lead to adequate matching of examinees.

Our findings, in general, appear to provide good news 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 zero, although the CAT-based statistics showed a slight inflation, particularly for Pool 3, in which DIF and difficulty were positively correlated. In general, Pool 3 DIF statistics were not quite as well-behaved as the Pool 2 statistics. The values of the DIF statistics for DIFless items in Pool 3 were not as close to zero and the detection rate for nominal C items was lower.

The factors that affected the size of the *MH D-DIF* and *STD P-DIF* statistics, in general, were the size of the true DIF, the item difficulty, and the interactions of item

difficulty and item discrimination, respectively, with the true DIF. Focal group distribution, item position, and sample size conditions had almost no effect.

A finding that was useful, although not directly relevant to CATs, was that in nonadaptive administration of 75 items, matching on the expected true score based on the MLE of ability led to essentially the same results as matching on number-right score. The similarity between these approaches, however, may be substantially less for shorter tests.

In most cases, examinee residual abilities for both the CAT and the 75-item nonadaptive test had medians close to zero within a population group and pool. The major exception was the focal N (-1, 1) group that received Pool 3 items; these examinees had a median residual of about -.1 for the CAT and -.06 for the nonadaptive test.[10] (The standard deviation of true ability was unity in each population group.) The differences between median residuals for the CAT and those for the nonadaptive test were not large relative to their standard errors; therefore, our findings did not support the conjecture that CATs might be more disadvantageous than nonadaptive tests for lower-achieving groups when DIF and difficulty were positively correlated.

Like the DIF statistics for CAT items, the pretest DIF statistics were well-behaved and had high correlations with true DIF. The pretest DIF statistics tended to be very slightly

---

[10]As noted earlier, both the DIF statistics and the examinee residuals tended to show larger departures from their target values in Pool 3, in which DIF and difficulty were positively correlated, than in Pool 2, in which they were uncorrelated. The interpretation of this result is not clear-cut, however. The Pool 3 data set was created because of the empirical finding that DIF *estimates* are sometimes positively correlated with item difficulty *estimates*. This does not imply, however, that the appropriate data-generating model is one in which the true (and ordinarily unknown) DIF and difficulty *parameters* are correlated, In short, there is no solid evidence for determining whether the Pool 2 or the Pool 3 data generating-model is more realistic.

smaller in magnitude than the DIF values for CAT items with the same item parameters. A more striking difference between the CAT and pretest results was that the standard errors of the *MH D-DIF* statistics tended to be larger for the pretest items than for the CAT items, further reducing the power to detect DIF. Possible reasons for the larger standard errors include the different method of constructing the matching variable, the greater sparseness of the data, and the occurrence of items with larger percents correct than in the CAT. Previous research has shown that all of these features can affect the size of *SE(MH D-DIF)*.

There are many questions that our study did not address. For example, because we used constant sample sizes in our simulation, we did not address the problem of insufficient item data that may arise when conducting DIF analyses of adaptively administered items (Miller, 1992). We did not consider methods for refining the DIF criterion by deleting DIF items and repeating the analysis, nor did we evaluate the effects of using different procedures, such as Bayesian methods, for estimating abilities or item parameters. Our conclusions apply to the case in which the data generation model and the estimation model are both based on the 3PL function. Also, our results may depend on our use of the expected true score for the item pool as our matching variable. Other types of scores may be of interest. For example, in some actual CATs, an expected true score is computed for a set of reference items that are not, in fact, included in the item pool. We did not consider CAT algorithms in which item selection is not determined solely by information, but is constrained by requirements concerning item type and content, nor did we examine the effects of complex starting algorithms used in some CATs to control the "exposure" of items. Finally, our study could not, of course, provide any data on the appropriateness of using item parameter estimates

obtained through paper-and-pencil or nonadaptive computer administration to estimate item information and examinee ability in a CAT setting. If administration mode (see Hetter, Segall, & Bloxom, 1992; Wainer & Mislevy, 1990) or item order and context (see Zwick, 1991) affect the functioning of items, CAT-based ability estimation and hence DIF estimation will be impaired.

6.1 Opportunities for further research and applications

The data files we have created will facilitate further research on CATS at a relatively low cost. First, since we generated responses to all items in all three pools, we can create new CATS for the examinees without repeating the step of generating examinee abilities and item responses. Second, with the 2 x 2 x K tables of probabilities we generated for each of the 3 (pools) x 71 (administered items per pool) = 213 CAT items, as well as for the pretest items, we can create expected tables for any target sample sizes and compute DIF statistics on these tables without further data generation. An additional application of our work may involve the expected percent of A, B, and C results that can be computed according to Appendix D. It is likely that this method could be successfully applied to any large data set, such as the SAT. For example, the method could be used to predict the likelihood that an item would be categorized as a C item for various combinations of reference and focal group sample sizes. Viewing an item's DIF status as probabilistic, rather than deterministic, may be a fruitful way of evaluating DIF results.

References

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 and 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.

Hetter, R. D., Segall, D. O., & Bloxom, B. (October, 1992). *Item calibration medium effect on CAT scores.* Presented at the annual conference of the Military Testing Association, San Diego.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty.* ETS Research Report 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 and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Holland, P. W., & Zwick, R. (1991). *A simulation study of some simple approaches to the study of DIF for CATs.* Internal memorandum, Educational Testing Service.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty.* Report 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., (1976). A broad-range tailored test of verbal ability. In C. L. Clark(Ed.). *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington, DC.

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

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

McHale, F., Dorans, N., Holland, P., & Petersen, N. (May 2, 1988). *Specifications for standardized percent correct and distractor analysis (IANA80 and IANA82).* Educational Testing Service technical memo.

Miller, T. R. (April, 1992). *Practical considerations for conducting studies of differential item functioning (DIF) in a CAT environment.* 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.

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. (August 3, 1992). *Field test of a computer-based general test.* Draft GRE Board Professional Report.

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

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (ed.), *Computerized adaptive testing: A primer*, pp. 65-102. 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.* 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., (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, **10**, 3, 10-16.

Zwick, R. (October, 1992). *Differential item functioning analysis for computer-Adaptive tests and other IRT-scored measures.* Presented at the annual meeting of the Military Testing Association, San Diego.

Zwick, R. in collaboration with Donoghue, J., Grima, A., Holland, P., Thayer, D., Thomas, N., & Wingersky, M. (April, 1992). *Differential item functioning analysis for new modes of assessment.* Presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Appendix A

Supplementary Tables

Table A-1

Item Information Table [a]

Item Position	Ability Level																				
	-2.0	-1.8	-1.6	-1.4	-1.2	-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
1	40	40	43	43	43	43	46	45	45	52	52	56	56	60	60	69	69	69	74	74	74
2	39	39	40	41	41	41	45	46	46	45	56	52	52	56	69	65	65	66	69	66	73
3	42	42	41	42	42	50	50	48	48	54	54	54	60	52	65	68	66	65	66	69	75
4	41	41	42	44	44	46	49	49	49	48	53	53	64	64	64	66	68	67	67	67	71
5	44	44	44	40	50	44	43	50	47	56	55	55	53	62	68	60	67	74	65	73	67
6	43	43	39	39	46	42	48	47	55	55	45	51	54	29	62	67	60	68	68	65	66
7	2	2	2	50	49	49	41	55	54	46	48	64	62	61	29	29	29	73	73	71	69
8	1	1	8	8	40	45	47	43	50	53	51	60	61	53	56	64	64	71	71	75	65
9	5	5	50	49	39	48	44	54	52	47	46	45	55	59	61	62	62	60	75	68	68
10	3	8	1	46	8	47	42	41	51	49	47	57	63	63	52	61	74	29	70	70	72
11	4	7	5	2	45	8	55	51	53	51	49	61	51	54	59	59	61	75	60	72	70
12	8	9	9	9	48	55	17	44	56	50	57	48	29	28	66	28	59	62	29	32	32
13	7	6	6	1	47	17	54	42	57	57	22	62	59	65	28	63	28	70	32	35	35
14	6	11	7	6	9	40	51	53	43	22	24	63	24	68	63	26	71	64	72	36	36
15	9	50	11	7	17	39	8	17	17	17	50	24	28	26	26	56	26	61	62	60	38
16	11	4	10	5	2	15	15	57	22	24	63	22	23	58	58	58	70	59	35	29	33
17	10	3	49	11	6	14	57	15	15	15	23	23	26	55	67	52	63	28	64	62	34
18	50	10	46	10	55	9	14	52	41	25	61	47	57	24	53	70	73	32	59	59	60
19	14	49	4	45	14	51	12	8	44	13	64	46	58	51	54	53	58	35	61	33	29
20	18	14	3	14	15	16	16	22	13	63	25	59	22	69	24	35	35	26	36	64	59
21	16	46	14	17	7	12	53	14	42	23	17	49	45	23	23	54	32	63	28	34	62
22	49	16	15	15	10	54	13	13	25	43	62	58	48	57	55	23	75	36	26	61	61
23	12	18	17	16	11	6	9	12	14	19	60	26	25	22	51	24	36	58	63	38	28
24	15	12	16	12	1	10	22	16	24	20	58	28	19	27	27	27	56	72	33	28	26
25	17	15	12	47	16	57	18	56	12	12	59	25	47	19	22	71	27	33	34	26	64

[a] For each ability level, the table lists the 25 most informative items, starting with the most informative.

Table A-2

Item Usage Table[a]

Item	Reference Group-All Conditions	Condition: Pool: Focal Group:	1, 2 1 N(-1,1)	3, 4 2 N(-1,1)	5, 6 3 N(-1,1)	7, 8 1 N(0,1)	9, 10 2 N(0,1)	11, 12 3 N(0,1)	13, 14 1 N(.5,1)	15, 16 2 N(.5,1)	17, 18 3 N(.5,1)
1	8415		26638	26642	28532	8418	8663	9677	3656	3810	4508
2	10250		29964	29734	31443	10416	10603	11721	4746	4826	5666
3	4990		19332	19850	21338	4966	5279	5934	1839	2021	2424
4	5225		19938	20315	21931	5222	5478	6220	1970	2125	2588
5	7345		24621	24906	26706	7340	7697	8666	3050	3253	3941
6	10275		29707	29433	31095	10356	10589	11769	4807	4895	5713
7	8404		26579	26583	28253	8487	8772	9796	3695	3847	4563
8	19293		41043	40882	41954	19516	19504	20879	10664	10709	11847
9	12535		32894	32564	34310	12633	12824	14245	6161	6238	7228
10	7667		24477	24644	26212	7759	8033	8990	3356	3538	4188
11	7203		23906	24113	25792	7221	7588	8491	3017	3188	3890
12	9069		20748	21130	21441	9231	9434	9888	5089	5252	5548
13	10935		10426	10818	9360	10827	11041	10284	8196	8256	7856
14	17942		36805	36787	37037	18082	18191	18872	10301	10247	10941
15	23574		36395	36222	35367	23656	23534	23551	15526	15326	15621
16	8914		22661	22864	23581	9069	9261	9986	4631	4845	5314
17	25597		31532	31297	29571	25659	25672	25369	17742	17818	17765
18	2935		13505	13730	15053	2951	3092	3563	969	1087	1397
19	1074		816	741	581	1123	1081	812	943	869	761
20	287		227	245	167	291	323	223	230	269	195
21	0		0	0	0	0	0	0	0	0	0
22	24377		17864	18203	15758	24171	24135	21594	20909	20806	18770
23	22085		10490	10242	9048	21708	21033	18605	23240	22332	19701
24	27931		14917	14722	13106	27603	27192	24379	28025	27735	24799
25	10870		8032	8112	6942	10646	10827	9342	8895	8937	7915
26	22280		6601	6360	6415	22126	21451	21285	31845	30907	30491
27	1928		593	435	463	1948	1571	1480	2509	2137	1960
28	24589		7334	7089	7236	24417	23928	24086	35473	34855	34882
29	27744		9426	9174	9401	27687	27097	27317	38502	37986	38044
30	0		0	0	0	0	0	0	0	0	0
31	0		0	0	0	0	0	0	0	0	0
32	8465		1339	1386	1655	8508	8448	9691	16459	16420	18150
33	3829		462	433	648	3960	3882	5007	9037	8953	10823
34	2847		299	290	463	2938	2795	3757	7032	6763	8580
35	11396		2112	2045	2308	11356	11102	12070	20600	19979	21452
36	7383		1104	1154	1366	7507	7446	8599	14970	14867	16508
37	0		0	0	0	0	0	0	0	0	0
38	2730		284	271	438	2828	2692	3638	6790	6587	8420

(continued)

Table A-2 (continued)

Item Usage Table

Item	Reference Group-All Conditions	Condition: Pool: Focal Group:	1, 2 1 N(-1,1)	3, 4 2 N(-1,1)	5, 6 3 N(-1,1)	7, 8 1 N(0,1)	9, 10 2 N(0,1)	11, 12 3 N(0,1)	13, 14 1 N(.5,1)	15, 16 2 N(.5,1)	17, 18 3 N(.5,1)
39	19334	39223	39185	40120	19494	20124	20632	11880	12460	12799	
40	20571		40364	40938	42047	20716	21894	22652	12769	16472	17793
41	25646		46415	46786	47717	25770	26564	27938	15769	16472	17793
42	25606		45831	46039	46678	25855	26632	27234	16361	16893	17540
43	31779		49802	50182	50584	31729	32733	33003	21753	22822	23081
44	23694		45110	45170	46119	23901	24164	25540	14101	14255	15397
45	40033		42937	42879	41633	40107	40544	38925	31901	32518	30943
46	35983		43356	43508	42125	36173	36825	35768	26671	27586	26863
47	31113		32538	32551	30491	31349	31685	30234	23621	24223	23212
48	34650		35446	35394	33380	34785	35075	33446	26965	27611	26306
49	35171		50533	50885	50344	35348	36165	35668	24583	25433	25238
50	32939		50796	51133	51105	33008	33891	33951	22047	22862	23100
51	38251		30310	30357	27931	38326	38059	35863	33472	33507	31468
52	32641		24523	24225	23170	32652	32185	30436	30972	30564	28902
53	30585		20814	21024	19625	30489	30465	28423	29083	28985	27156
54	32646		24921	24822	23143	32672	32225	30699	30206	30045	28272
55	41787		35621	35316	33120	42004	41662	39726	36309	36405	34397
56	32613		22866	21734	21233	32080	31225	29861	31956	31447	29464
57	32045		24659	24797	22365	31923	31696	29237	27496	27635	25359
58	19429		6496	6200	5858	19178	18332	17213	24742	23573	21857
59	28223		9289	9174	8949	28143	27664	27318	39078	38618	38151
60	39658		21441	20987	21139	39519	38481	38372	47463	46967	46626
61	33879		13139	13093	12421	33652	33374	32432	43865	43624	42603
62	32991		12851	12699	12327	32881	32398	31702	43098	42677	41898
63	32944		14255	14336	13131	32608	32686	30216	38207	38323	35203
64	34882		16068	15762	15268	34676	33944	32789	41402	40887	38994
65	27212		10877	10870	11339	27240	27072	27659	37368	37292	37581
66	22325		8053	8168	8702	22288	22441	23365	32437	32572	33366
67	18876		5402	5434	6013	18811	18719	19837	29188	28878	30105
68	24352		8392	8601	8970	24399	24422	25037	34944	34917	35296
69	33521		18075	17813	18501	33567	32792	33746	42007	41432	42058
70	12844		2530	2468	2729	12716	12494	13296	22438	21762	23017
71	11833		2669	2825	3265	11751	12076	13353	20539	20832	22495
72	5975		832	872	1136	6026	6112	7445	12644	12668	14753
73	20677		9769	9729	10418	20619	20552	21891	28806	28852	30248
74	24136		11893	11940	12792	24028	24067	25550	32471	32558	34051
75	18773		8833	8688	9141	18891	18303	19757	26514	26371	27644

[a]Each entry represents the number of examinees who received the indicated item out of a total of 60,000 examinees.

Table A-3

Item Use By Ability Level for Reference Group

Proportion of Examinees in Each Interval Who Received Each Item[a]

Item	<2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.01	0.03	0.07	0.12	0.17	0.20	0.18	0.12	0.07	0.03	0.00
1	0.99	0.96	0.77	0.35	0.05	0.00	0. 00	0.00	0.00	0.00	0.00
2	1.00	0.98	0.86	0.48	0.09	.0.01	0. 00	0.00	0.00	0.00	0.00
3	0.98	0.87	0.50	0.12	0.01	0.00	0. 00	0.00	0.00	0.00	0.00
4	0.98	0.88	0.52	0.13	0.01	0.00	0. 00	0.00	0.00	0.00	0.00
5	0.99	0.94	0.71	0.27	0.03	0.00	0. 00	0.00	0.00	0.00	0.00
6	0.97	0.96	0.84	0.50	0.10	0.01	0. 00	0.00	0.00	0.00	0.00
7	0.98	0.95	0.77	0.36	0.05	0.00	0. 00	0.00	0.00	0.00	0.00
8	0.99	0.99	0.97	0.89	0.52	0.11	0. 01	0.00	0.00	0.00	0.00
9	0.96	0.97	0.90	0.64	0.18	0.02	0. 00	0.00	0.00	0.00	0.00
10	0.93	0.90	0.69	0.32	0.05	0.00	0. 00	0.00	0.00	0.00	0.00
11	0.95	0.91	0.68	0.27	0.03	0.00	0. 00	0.00	0.00	0.00	0.00
12	0.74	0.57	0.32	0.37	0.29	0.06	0. 00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.04	0.23	0.50	0.30	0. 05	0.00	0.00	0.00	0.00
14	0.88	0.87	0.82	0.79	0.52	0.13	0. 01	0.00	0.00	0.00	0.00
15	0.67	0.57	0.62	0.80	0.80	0.41	0. 07	0.00	0.00	0.00	0.00
16	0.79	0.70	0.46	0.39	0.21	0.03	0. 00	0.00	0.00	0.00	0.00
17	0.03	0.18	0.58	0.88	0.88	0.53	0. 12	0.01	0.00	0.00	0.00
18	0.82	0.65	0.25	0.03	0.00	0.00	0. 00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.01	0.03	0.04	0. 01	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.01	0.01	0. 00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.03	0.25	0.69	0.80	0. 47	0.10	0.01	0.00	0.00
23	0.00	0.00	0.00	0.02	0.22	0.62	0. 76	0.50	0.12	0.01	0.00
24	0.00	0.00	0.01	0.07	0.40	0.82	0. 87	0.51	0.11	0.01	0.00
25	0.00	0.00	0.01	0.08	0.36	0.42	0. 14	0.01	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.03	0.24	0. 68	0.93	0.83	0.69	0.90
27	0.00	0.00	0.00	0.00	0.00	0.01	0. 07	0.12	0.05	0.00	0.00
28	0.00	0.00	0.00	0.00	0.04	0.26	0. 73	0.97	0.97	0.96	0.99
29	0.00	0.00	0.01	0.02	0.09	0.37	0. 81	0.99	1.00	1.00	1.00
32	0.00	0.00	0.00	0.00	0.00	0.00	0. 05	0.33	0.80	0.98	1.00

(continued)

Table A-3 (Continued)

Item Use by Ability Level for Reference Group

Proportion of Examinees in Each Interal Who Received Each Item

Item	<2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.01	0.03	0.07	0.12	0.17	0.20	0.18	0.12	0.07	0.03	0.00
33	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.05	0.35	0.81	0.99
34	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.02	0.22	0.68	0.97
35	0.00	0.00	0.00	0.00	0.00	0.01	0. 13	0.54	0.90	0.98	1.00
36	0.00	0.00	0.00	0.00	0.00	0.00	0. 03	0.26	0.72	0.95	1.00
38	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.02	0.20	0.67	0.96
39	1.00	0.99	0.95	0.76	0.40	0.18	0. 10	0.05	0.03	0.01	0.00
40	1.00	0.99	0.96	0.79	0.44	0.21	0. 11	0.05	0.02	0.01	0.00
41	1.00	1.00	0.99	0.97	0.76	0.32	0. 06	0.01	0.00	0.00	0.00
42	1.00	1.00	0.99	0.95	0.71	0.30	0. 11	0.05	0.02	0.01	0.00
43	1.00	1.00	1.00	0.98	0.86	0.52	0. 23	0.10	0.05	0.03	0.01
44	1.00	1.00	0.99	0.95	0.71	0.25	0. 03	0.00	0.00	0.00	0.00
45	0.28	0.41	0.73	0.96	0.99	0.91	0. 60	0.22	0.08	0.04	0.01
46	0.24	0.52	0.87	0.98	0.97	0.80	0. 39	0.10	0.03	0.01	0.00
47	0.02	0.11	0.44	0.84	0.96	0.78	0. 34	0.05	0.00	0.00	0.00
48	0.06	0.16	0.51	0.87	0.98	0.86	0. 47	0.10	0.01	0.00	0.00
49	0.85	0.91	0.95	0.99	0.96	0. /4	0. 30	0.04	0.01	0.00	0.00
50	0.95	0.98	0.99	0.99	0.94	0.65	0. 19	0.02	0.00	0.00	0.00
51	0.00	0.03	0.21	0.64	0.94	0.97	0. 81	0.36	0.06	0.01	0.00
52	0.38	0.38	0.47	0.65	0.90	0.99	0. 98	0.82	0.51	0.31	0.28
53	0.29	0.31	0.36	0.58	0.90	0.99	0. 97	0.78	0.43	0.27	0.25
54	0.32	0.32	0.46	0.74	0.96	1.00	0. 96	0.76	0.42	0.28	0.21
55	0.03	0.09	0.38	0.81	0.98	0.99	0. 85	0.42	0.08	0.01	0.00
56	0.38	0.36	0.45	0.57	0.84	0.98	0. 99	0.90	0.59	0.37	0.28
60	0.11	0.13	0.17	0.24	0.43	0.74	0. 95	1.00	1.00	1.00	1.00
61	0.01	0.00	0.01	0.03	0.20	0.66	0. 95	1.00	0.99	0.98	0.99
62	0.01	0.01	0.02	0.04	0.20	0.60	0. 93	1.00	1.00	1.00	1.00
63	0.00	0.00	0.01	0.04	0.29	0.75	0. 97	0.96	0.74	0.29	0.03
64	0.03	0.03	0.06	0.10	0.30	0.71	0. 96	1.00	0.94	0.62	0.12
65	0.03	0.03	0.05	0.07	0.15	0.34	0. 71	0.96	1.00	1.00	1.00
66	0.03	0.02	0.03	0.05	0.09	0.21	0. 52	0.88	0.99	1.00	1.00
67	0.01	0.01	0.01	0.02	0.04	0.12	0. 41	0.84	0.99	1.00	1.00

(continued)

Table A-3 (Continued)

Item Use by Ability Level for Reference Group

Proportion of Examinees in Each Interal Who Received Each Item

Item	<2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.01	0.03	0.07	0.12	0.17	0.20	0.18	0.12	0.07	0.03	0.00
68	0.01	0.02	0.03	0.04	0.09	0.26	0. 64	0.94	1.00	1.00	1.00
69	0.12	0.12	0.16	0.21	0.32	0.52	0. 78	0.96	1.00	1.00	1.00
70	0.00	0.00	0.00	0.00	0.00	0.02	0. 18	0.63	0.94	1.00	1.00
71	0.00	0.00	0.01	0.01	0.01	0.04	0. 15	0.51	0.88	0.99	1.00
72	0.00	0.00	0.00	0.00	0.00	0.00	0. 01	0.16	0.60	0.93	1.00
73	0.06	0.06	0.09	0.12	0.17	0.25	0. 38	0.64	0.90	0.99	1.00
74	0.09	0.08	0.11	0.14	0.20	0.31	0. 48	0.75	0.95	1.00	1.00
75	0.05	0.07	0.08	0.10	0.15	0.22	0. 33	0.56	0.87	0.99	1.00

*For all examinees, the first item was selected randomly from among items 52, 53, 54, and 56. If the first response was correct, the next item was selected randomly from among items 73, 74 and 75. If it was incorrect, the next item was selected randomly from among items 39, 40 and 42.

Table A-4

Item Use by Ability Level for N(-1,1) Focal Group, Pool 3

Proportion of Examinees in Each Interval Who Received Each Item[a]

Item	<-2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.11	0.12	0.17	0.20	0.17	0.12	0.07	0.03	0.01	0.00	0.00
1	0.99	0.98	0.84	0.46	0.09	0.01	0. 00	0.00	0.00	0.00	0.00
2	0.99	0.99	0.90	0.58	0.16	0.01	0. 00	0.00	0.00	0.00	0.00
3	0.98	0.90	0.60	0.19	0.02	0.00	0. 00	0.00	0.00	0.00	0.00
4	0.98	0.91	0.63	0.21	0.02	0.00	0. 00	0.00	0.00	0.00	0.00
5	0.99	0.96	0.79	0.38	0.07	0.00	0. 00	0.00	0.00	0.00	0.00
6	0.97	0.96	0.88	0.59	0.17	0.01	0. 00	0.00	0.00	0.00	0.00
7	0.98	0.96	0.83	0.46	0.10	0.00	0. 00	0.00	0.00	0.00	0.00
8	0.99	0.99	0.98	0.91	0.60	0.16	0. 01	0.00	0.00	0.00	0.00
9	0.97	0.96	0.92	0.72	0.28	0.03	0. 00	0.00	0.00	0.00	0.00
10	0.94	0.90	0.75	0.42	0.09	0.01	0. 00	0.00	0.00	0.00	0.00
11	0.95	0.92	0.76	0.38	0.07	0.00	0. 00	0.00	0.00	0.00	0.00
12	0.80	0.63	0.37	0.34	0.32	0.09	0. 00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.02	0.17	0.43	0.32	0. 06	0.00	0.00	0.00	0.00
14	0.91	0.86	0.81	0.78	0.56	0.18	0. 01	0.00	0.00	0.00	0.00
15	0.71	0.58	0.56	0.74	0.78	0.46	0. 10	0.00	0.00	0.00	0.00
16	0.82	0.72	0.51	0.39	0.26	0.06	0. 00	0.00	0.00	0.00	0.00
17	0.01	0.12	0.46	0.83	0.87	0.58	0. 15	0.01	0.00	0.00	0.00
18	0.86	0.71	0.36	0.06	0.00	0.00	0. 00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.02	0.04	0. 01	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.01	0.01	0. 00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.02	0.17	0.57	0.76	0. 46	0.08	0.00	0.00	0.00
23	0.00	0.00	0.00	0.01	0.15	0.51	0. 71	0.41	0.08	0.01	0.00
24	0.00	0.00	0.00	0.04	0.31	0.73	0. 82	0.45	0.07	0.01	0.00
25	0.00	0.00	0.00	0.05	0.27	0.40	0. 15	0.01	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.03	0.20	0. 64	0.90	0.76	0.78	0.92
27	0.00	0.00	0.00	0.00	0.00	0.01	0. 06	0.09	0.03	0.01	0.00
28	0.00	0.00	0.00	0.00	0.04	0.24	0. 69	0.96	0.96	0.97	1.00
29	0.01	0.01	0.01	0.03	0.09	0.34	0. 77	0.98	1.00	0.98	1.00
32	0.00	0.00	0.00	0.00	0.00	0.01	0. 07	0.39	0.86	0.99	1.00
33	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.09	0.48	0.92	1.00

(continued)

Table A-4 (continued)

Item Use by Ability Level for N(-1,1) Focal Group, Pool 3

Proportion of Examinees in Each Interval Who Received Each Item

Item	<-2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.11	0.12	0.17	0.20	0.17	0.12	0.07	0.03	0.01	0.00	0.00
34	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.04	0.35	0.83	1.00
35	0.00	0.00	0.00	0.00	0.00	0.01	0. 14	0.57	0.92	1.00	1.00
36	0.00	0.00	0.00	0.00	0.00	0.00	0. 04	0.31	0.79	0.98	1.00
38	0.00	0.00	0.00	0.00	0.00	0.00	0. 00	0.04	0.33	0.80	0.97
39	1.00	0.99	0.96	0.81	0.47	0.21	0. 11	0.06	0.03	0.01	0.00
40	1.00	1.00	0.97	0.86	0.54	0.27	0. 12	0.05	0.04	0.01	0.00
41	1.00	1.00	1.00	0.98	0.82	0.43	0. 12	0.03	0.01	0.00	0.00
42	1.00	1.00	0.99	0.96	0.76	0.38	0. 12	0.06	0.02	0.01	0.00
43	1.00	1.00	1.00	0.99	0.89	0.60	0. 26	0.11	0.07	0.02	0.00
44	1.00	1.00	0.99	0.97	0.77	0.34	0. 05	0.00	0.00	0.00	0.00
45	0.22	0.37	0.66	0.93	0.99	0.91	0. 59	0.22	0.09	0.02	0.00
46	0.19	0.44	0.79	0.97	0.97	0.81	0. 41	0.11	0.03	0.01	0.00
47	0.02	0.09	0.34	0.75	0.94	0.79	0. 37	0.05	0.01	0.00	0.00
48	0.04	0.13	0.40	0.80	0.96	0.86	0. 48	0.11	0.02	0.00	0.00
49	0.86	0.88	0.94	0.98	0.96	0.77	0. 34	0.05	0.01	0.00	0.00
50	0.96	0.97	0.98	0.99	0.95	0.70	0. 26	0.03	0.01	0.00	0.00
51	0.01	0.02	0.14	0.52	0.89	0.96	0. 79	0.32	0.04	0.00	0.00
52	0.36	0.39	0.45	0.60	0.85	0.98	0. 97	0.78	0.41	0.38	0.28
53	0.27	0.28	0.32	0.51	0.84	0.98	0. 96	0.74	0.35	0.20	0.28
54	0.29	0.33	0.41	0.66	0.92	0.99	0. 95	0.71	0.41	0.30	0.26
55	0.02	0.07	0.28	0.70	0.96	0.98	0. 84	0.37	0.06	0.00	0.00
56	0.37	0.38	0.42	0.52	0.75	0.96	0. 98	0.86	0.50	0.27	0.18
57	0.00	0.01	0.07	0.35	0.77	0.89	0. 63	0.17	0.01	0.00	0.00
58	0.00	0.00	0.00	0.00	0.03	0.21	0. 61	0.81	0.44	0.06	0.00
59	0.00	0.00	0.00	0.01	0.07	0.35	0. 81	0.98	0.99	0.98	1.00
60	0.11	0.12	0.17	0.23	0.39	0.68	0. 93	1.00	1.00	1.00	1.00
61	0.00	0.00	0.01	0.03	0.17	0.57	0. 93	0.99	0.99	0.97	1.00
62	0.01	0.01	0.02	0.04	0.17	0.52	0. 90	1.00	1.00	0.98	1.00
63	0.00	0.00	0.00	0.03	0.22	0.66	0. 95	0.94	0.60	0.17	0.00
64	0.03	0.03	0.06	0.10	0.26	0.62	0. 93	0.99	0.86	0.41	0.10
65	0.03	0.04	0.05	0.08	0.16	0.35	0. 69	0.96	0.99	1.00	1.00
66	0.02	0.03	0.04	0.06	0.11	0.24	0. 54	0.90	0.99	1.00	1.00

(continued)

Table A-4 (continued)

Item Use by Ability Level for N(-1,1) Focal Group, Pool 3

Proportion of Examinees in Each Interval Who Received Each Item

Item	<-2.25	-2.25 to -1.75	-1.75 to -1.25	-1.25 to -0.75	-0.75 to -0.25	-0.25 to 0.25	0.25 to 0.75	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	>2.25
	Distribution of Examinees Across Ability Intervals										
	0.11	0.12	0.17	0.20	0.17	0.12	0.07	0.03	0.01	0.00	0.00
67	0.01	0.01	0.02	0.02	0.05	0.14	0. 44	0.85	0.99	1.00	1.00
68	0.02	0.02	0.03	0.05	0.11	0.27	0. 63	0.94	0.99	1.00	1.00
69	0.11	0.13	0.17	0.22	0.33	0.51	0. 75	0.96	1.00	1.00	1.00
70	0.00	0.00	0.00	0.00	0.00	0.02	0. 18	0.64	0.95	1.00	1.00
71	0.00	0.00	0.01	0.01	0.02	0.05	0. 18	0.57	0.92	0.99	1.00
72	0.00	0.00	0.00	0.00	0.00	0.00	0. 03	0.22	0.73	0.96	1.00
73	0.06	0.07	0.09	0.12	0.18	0.25	0. 39	0.68	0.94	0.99	1.00
74	0.08	0.09	0.11	0.15	0.22	0.33	0. 51	0.80	0.97	1.00	1.00
75	0.06	0.06	0.08	0.11	0.15	0.22	0. 34	0.61	0.91	0.99	1.00

[1]For all examinees, the first item was selected randomly from among items 52, 53, 54, and 56. If the first response was correct, the next item was selected randomly from among items 73, 74 and 75. If it was incorrect, the next item was selected randomly from among items 39, 40 and 42.

Table A-5

Frequency of Item Parameter Combinations in Pool 1 (75 Items)

	ln a	
b	-.30	0.00
-1.95	5	2
-1.30	6	4
-0.65	7	6
0.00	7	7
0.65	6	7
1.30	5	6
1.95	2	5

Table A-6

Frequency of Item Parameter Combinations for Pool 2 (75 Items)

ln a = -0.30	*d*					
b	-0.50	-0.25	0.	0.25	0.50	Marginal
-1.95	0	2	2	1	0	5
-1.30	0	2	2	2	0	6
-0.65	0	2	3	2	0	7
0.00	1	1	2	2	1	7
0.65	0	1	2	2	1	6
1.30	1	1	2	1	0	5
1.95	0	0	1	1	0	2
Marginal	2	9	14	11	2	38

ln a = 0.00	*d*					
b	-0.50	-0.25	0.	0.25	0.50	Marginal
-1.95	0	1	1	0	0	2
-1.30	0	1	2	1	0	4
-0.65	0	2	2	1	1	6
0.00	1	2	2	1	1	7
0.65	0	2	3	2	0	7
1.30	0	2	2	2	0	6
1.95	0	1	2	2	0	5
Marginal	1	11	14	9	2	37

Table A-7

Frequency of Item Parameter Combinations in Pool 3 (75 Items)

ln a = -0.30	*d*					
b	-0.50	-0.25	0.0	0.25	0.50	Marginal
-1.95	1	2	1	1	0	5
-1.30	1	2	2	1	0	6
-0.65	0	2	3	2	0	7
0.00	0	2	3	2	0	7
0.65	0	1	2	2	1	6
1.30	0	1	2	1	1	5
1.95	0	0	1	1	0	2
Marginal	2	10	14	10	2	38

ln a = 0.00	*d*					
b	-0.50	-0.25	0.0	0.25	0.50	Marginal
-1.95	0	1	1	0	0	2
-1.30	0	1	2	1	0	4
-0.65	1	2	2	1	0	6
0.00	0	2	3	2	0	7
0.65	0	2	3	2	0	7
1.30	0	1	2	2	1	6
1.95	0	1	1	2	1	5
Marginal	1	10	14	10	2	37

Table A-8

a, *b*, and *d* Parameters in Pool 2 and Pool 3

	Pools 1, 2, & 3		Pool 2			Pool 3		
Item	*a*	*b*	*b-d*	*d*	*ad*	*b-d*	*d*	*ad*
1	0.74	-1.95	-1.60	-0.35	-0.26	-1.25	-0.70	-0.52
2	0.74	-1.95	-1.60	-0.35	-0.26	-1.60	-0.35	-0.26
3	0.74	-1.95	-1.95	0.00	0.00	-1.60	-0.35	-0.26
4	0.74	-1.95	-1.95	0.00	0.00	-1.95	0.00	0.00
5	0.74	-1.95	-2.30	0.35	0.26	-2.30	0.35	0.26
6	0.74	-1.30	-0.95	-0.35	-0.26	-0.60	-0.70	-0.52
7	0.74	-1.30	-0.95	-0.35	-0.26	-0.95	-0.35	-0.26
8	0.74	-1.30	-1.30	0.00	0.00	-0.95	-0.35	-0.26
9	0.74	-1.30	-1.30	0.00	0.00	-1.30	0.00	0.00
10	0.74	-1.30	-1.65	0.35	0.26	-1.30	0.00	0.00
11	0.74	-1.30	-1.65	0.35	0.26	-1.65	0.35	0.26
12	0.74	-0.65	-0.30	-0.35	-0.26	-0.30	-0.35	-0.26
13	0.74	-0.65	-0.30	-0.35	-0.26	-0.30	-0.35	-0.26
14	0.74	-0.65	-0.65	0.00	0.00	-0.65	0.00	0.00
15	0.74	-0.65	-0.65	0.00	0.00	-0.65	0.00	0.00
16	0.74	-0.65	-0.65	0.00	0.00	-0.65	0.00	0.00
17	0.74	-0.65	-1.00	0.35	0.26	-1.00	0.35	0.26
18	0.74	-0.65	-1.00	0.35	0.26	-1.00	0.35	0.26
19	0.74	0.00	0.70	-0.70	-0.52	0.35	-0.35	-0.26
20	0.74	0.00	0.35	-0.35	-0.26	0.35	-0.35	-0.26
21	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.74	0.00	-0.35	0.35	0.26	0.00	0.00	0.00
24	0.74	0.00	-0.35	0.35	0.26	-0.35	0.35	0.26
25	0.74	0.00	-0.70	0.70	0.52	-0.35	0.35	0.26
26	0.74	0.65	1.00	-0.35	-0.26	1.00	-0.35	-0.26
27	0.74	0.65	0.65	0.00	0.00	0.65	0.00	0.00
28	0.74	0.65	0.65	0.00	0.00	0.65	0.00	0.00
29	0.74	0.65	0.30	0.35	0.26	0.30	0.35	0.26
30	0.74	0.65	0.30	0.35	0.26	0.30	0.35	0.26
31	0.74	0.65	-0.05	0.70	0.52	-0.05	0.70	0.52
32	0.74	1.30	2.00	-0.70	-0.52	1.65	-0.35	-0.26
33	0.74	1.30	1.65	-0.35	-0.26	1.30	0.00	0.00
34	0.74	1.30	1.30	0.00	0.00	1.30	0.00	0.00
35	0.74	1.30	1.30	0.00	0.00	0.95	0.35	0.26
36	0.74	1.30	0.95	0.35	0.26	0.60	0.70	0.52
37	0.74	1.95	1.95	0.00	0.00	1.95	0.00	0.00
38	0.74	1.95	1.60	0.35	0.26	1.60	0.35	0.26
39	1.00	-1.95	-1.60	-0.35	-0.35	-1.60	-0.35	-0.35
40	1.00	-1.95	-1.95	0.00	0.00	-1.95	0.00	0.00
41	1.00	-1.30	-0.95	-0.35	-0.35	-0.95	-0.35	-0.35
42	1.00	-1.30	-1.30	0.00	0.00	-1.30	0.00	0.00

(continued)

Table A-8 (continued)

a, *b*, and *d* Parameters in Pool 2 and Pool 3

	Pools 1, 2, & 3		Pool 2			Pool 3		
Item	*a*	*b*	*b-d*	*d*	*ad*	*b-d*	*d*	*ad*
43	1.00	-1.30	-1.30	0.00	0.00	-1.30	0.00	0.00
44	1.00	-1.30	-1.65	0.35	0.35	-1.65	0.35	0.35
45	1.00	-0.65	-0.30	-0.35	-0.35	0.05	-0.70	-0.70
46	1.00	-0.65	-0.30	-0.35	-0.35	-0.30	-0.35	-0.35
47	1.00	-0.65	-0.65	0.00	0.00	-0.30	-0.35	-0.35
48	1.00	-0.65	-0.65	0.00	0.00	-0.65	0.00	0.00
49	1.00	-0.65	-1.00	0.35	0.35	-0.65	0.00	0.00
50	1.00	-0.65	-1.35	0.70	0.70	-1.00	0.35	0.35
51	1.00	0.00	0.70	-0.70	-0.70	0.35	-0.35	-0.35
52	1.00	0.00	0.35	-0.35	-0.35	0.35	-0.35	-0.35
53	1.00	0.00	0.35	-0.35	-0.35	0.00	0.00	0.00
54	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
55	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
56	1.00	0.00	-0.35	0.35	0.35	-0.35	0.35	0.35
57	1.00	0.00	-0.70	0.70	0.70	-0.35	0.35	0.35
58	1.00	0.65	1.00	-0.35	-0.35	1.00	-0.35	-0.35
59	1.00	0.65	1.00	-0.35	-0.35	1.00	-0.35	-0.35
60	1.00	0.65	0.65	0.00	0.00	0.65	0.00	0.00
61	1.00	0.65	0.65	0.00	0.00	0.65	0.00	0.00
62	1.00	0.65	0.65	0.00	0.00	0.65	0.00	0.00
63	1.00	0.65	0.30	0.35	0.35	0.30	0.35	0.35
64	1.00	0.65	0.30	0.35	0.35	0.30	0.35	0.35
65	1.00	1.30	1.65	-0.35	-0.35	1.65	-0.35	-0.35
66	1.00	1.30	1.65	-0.35	-0.35	1.30	0.00	0.00
67	1.00	1.30	1.30	0.00	0.00	1.30	0.00	0.00
68	1.00	1.30	1.30	0.00	0.00	0.95	0.35	0.35
69	1.00	1.30	0.95	0.35	0.35	0.95	0.35	0.35
70	1.00	1.30	0.95	0.35	0.35	0.60	0.70	0.70
71	1.00	1.95	2.30	-0.35	-0.35	2.30	-0.35	-0.35
72	1.00	1.95	1.95	0.00	0.00	1.95	0.00	0.00
73	1.00	1.95	1.95	0.00	0.00	1.60	0.35	0.35
74	1.00	1.95	1.60	0.35	0.35	1.60	0.35	0.35
75	1.00	1.95	1.60	0.35	0.35	1.25	0.70	0.70

Table A-9

Pretest Item Parameters

Item	*a*	*b*	*b-d*	*d*	*c*
1	1.00	-1.30	-0.60	-0.70	0.15
2	1.00	-1.30	-0.95	-0.35	0.15
3	1.00	-1.30	-1.30	0.00	0.15
4	1.00	-1.30	-1.65	0.35	0.15
5	1.00	-1.30	-2.00	0.70	0.15
6	1.00	0.00	0.70	-0.70	0.15
7	1.00	0.00	0.35	-0.35	0.15
8	1.00	0.00	0.00	0.00	0.15
9	1.00	0.00	-0.35	0.35	0.15
10	1.00	0.00	-0.70	0.70	0.15
11	1.00	1.30	2.00	-0.70	0.15
12	1.00	1.30	1.65	-0.35	0.15
13	1.00	1.30	1.30	0.00	0.15
14	1.00	1.30	0.95	0.35	0.15
15	1.00	1.30	0.60	0.70	0.15

Table A-10

True and Estimated Item Parameters for Pools 1, 2, and 3

	True Item Parameters			Estimated Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
1	0.74	-1.95	0 15	0.73	-2.06	0.14
2	0.74	-1.95	0.15	0.75	-1.91	0.14
3	0.74	-1.95	0.15	0.64	-2.31	0.14
4	0.74	-1.95	0.15	0.64	-2.18	0.14
5	0.74	-1.95	0.15	0.73	-2.08	0.14
6	0.74	-1.30	0.15	0.70	-1.45	0.14
7	0.74	-1.30	0.15	0.69	-1.53	0.14
8	0.74	-1.30	0.15	0.80	-1.25	0.14
9	0.74	-1.30	0.15	0.72	-1.36	0.14
10	0.74	-1.30	0.15	0.68	-1.35	0.14
11	0.74	-1.30	0.15	0.68	-1.48	0.14
12	0.74	-0.65	0.15	0.73	-0.69	0.14
13	0.74	-0.65	0.15	0.81	-0.47	0.21
14	0.74	-0.65	0.15	0.75	-0.81	0.14
15	0.74	-0.65	0.15	0.79	-0.66	0.14
16	0.74	-0.65	0.15	0.73	-0.73	0.14
17	0.74	-0.65	0.15	0.82	-0.63	0.14
18	0.74	-0.65	0.15	0.69	-0.76	0.14
19	0.74	0.00	0.15	0.79	0.07	0.18
20	0.74	0.00	0.15	0.67	-0.23	0.07
21	0.74	0.00	0.15	0.68	-0.06	0.12
22	0.74	0.00	0.15	0.83	-0.07	0.12
23	0.74	0.00	0.15	0.92	0.18	0.23
24	0.74	0.00	0.15	0.91	0.11	0.19
25	0.74	0.00	0.15	0.78	-0.07	0.13
26	0.74	0.65	0.15	0.89	0.62	0.15
27	0.74	0.65	0.15	0.73	0.56	0.12
28	0.74	0.65	0.15	0.93	0.64	0.18
29	0.74	0.65	0.15	1.08	0.74	0.22
30	0.74	0.65	0.15	0.61	0.47	0.06
31	0.74	0.65	0.15	0.59	0.42	0.04
32	0.74	1.30	0.15	0.83	1.41	0.19
33	0.74	1.30	0.15	0.69	1.22	0.14
34	0.74	1.30	0.15	0.66	1.31	0.10
35	0.74	1.30	0.15	0.79	1.22	0.16
36	0.74	1.30	0.15	0.75	1.29	0.15
37	0.74	1.95	0.15	0.54	2.28	0.13
38	0.74	1.95	0.15	0.6	1.94	0.13
39	1.00	-1.95	0.15	0.97	-2.02	0.14
40	1.00	-1.95	0.15	1.01	-2.01	0.14
41	1.00	-1.30	0.15	1.03	-1.37	0.14
42	1.00	-1.30	0.15	1.00	-1.44	0.14

(continued)

Table A-10 (continued)

True and Estimated Item Parameters for Pools 1, 2, and 3

	True Item Parameters			Estimated Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
43	1.00	-1.30	0.15	1.07	-1.30	0.14
44	1.00	-1.30	0.15	1.00	-1.40	0.14
45	1.00	-0.65	0.15	1.11	-0.58	0.17
46	1.00	-0.65	0.15	1.05	-0.69	0.14
47	1.00	-0.65	0.15	1.07	-0.59	0.22
48	1.00	-0.65	0.15	1.10	-0.55	0.20
49	1.00	-0.65	0.15	1.00	-0.70	0.12
50	1.00	-0.65	0.15	0.97	-0.83	0.11
51	1.00	0.00	0.15	0.92	-0.15	0.10
52	1.00	0.00	0.15	1.20	0.04	0.19
53	1.00	0.00	0.15	1.02	-0.03	0.14
54	1.00	0.00	0.15	0.98	-0.08	0.09
55	1.00	0.00	0.15	0.92	-0.20	0.05
56	1.00	0.00	0.15	1.20	0.11	0.19
57	1.00	0.00	0.15	0.86	-0.16	0.10
58	1.00	0.65	0.15	0.82	0.52	0.10
59	1.00	0.65	0.15	0.90	0.65	0.12
60	1.00	0.65	0.15	1.13	0.69	0.16
61	1.00	0.65	0.15	0.95	0.59	0.12
62	1.00	0.65	0.15	1.02	0.64	0.15
63	1.00	0.65	0.15	0.86	0.49	0.09
64	1.00	0.65	0.15	1.05	0.60	0.16
65	1.00	1.30	0.15	1.26	1.20	0.15
66	1.00	1.30	0.15	1.33	1.33	0.19
67	1.00	1.30	0.15	1.27	1.38	0.19
68	1.00	1.30	0.15	1.15	1.17	0.14
69	1.00	1.30	0.15	1.42	1.23	0.18
70	1.00	1.30	0.15	0.87	1.44	0.15
71	1.00	1.95	0.15	1.06	1.81	0.12
72	1.00	1.95	0.15	0.89	2.04	0.16
73	1.00	1.95	0.15	1.17	1.84	0.16
74	1.00	1.95	0.15	1.53	1.81	0.17
75	1.00	1.95	0.15	1.09	1.92	0.16

Appendix B

Variance of the ET Estimator of *MH D-DIF*[a]

[a]This variance was derived by Charles Lewis based on the work of Phillips and Holland (1987).

Let $\hat{\alpha}^{*}_{MH}$ be the MH odds ratio computed on the adjusted table frequencies, and *MH D-DIF** be the ET estimate of *MH D-DIF* based on the target sample sizes n_R^{*} and n_F^{*}. Then, based on the results in equations 4 - 6,

$$SE_{ET}(MH\ D\text{-}DIF^{*}) = 2.35\sqrt{Var\left(ln\left(\hat{\alpha}^{*}_{MH}\right)\right)},$$

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

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

where

$$U_k^{*} = (A_k\ D_k) + \hat{\alpha}^{*}_{MH}\ (B_k\ C_k)$$
$$V_k^{*} = (A_k + D_k) + \hat{\alpha}^{*}_{MH}\ (B_k + C_k),$$

and $T_k^{*} = \frac{n_{Rk}\ n_R^{*}}{n_R} + \frac{n_{Fk}\ n_F^{*}}{n_F}$.

Appendix C

Investigation of the ET Estimation Procedure

To check on the validity of the ET estimation procedure, we compared the results obtained using the ET method to those that we would have obtained using a more standard simulation procedure. We wanted to compare estimation procedures in the worst possible case; therefore, we chose Condition 5, which has the less stable sample size condition (n_R = 900, n_F = 100), the largest between-group ability difference (the focal N(-1, 1) population), and the most complex DIF structure (Pool 3). It would have been extremely expensive to conduct this validity check with the CAT data, in which each record includes different subsets of items. Therefore, we used the data from the 15 pretest items. (Note that the identity of the item pool was therefore relevant only to the matching variable; the DIF structure within the pretest items was the same in all conditions; see section 2.3.5). Because we had already generated data for 60,000 examinees in each group, we could use existing data to create 66 independent replications of the DIF analysis. (This number is the result of dividing 60,000 by 900 and then rounding down to the next lowest integer.) Table C-1 gives a comparison of the *MH D-DIF* results obtained from the 66 replications to the ET estimates reported in our study. The *STD P-DIF* findings yielded a similar picture of the agreement between the two estimation procedures. It is important to keep in mind that the two sets of results in Table C-1 are alternative estimates of unknown parameters; neither set can be regarded as the criterion. The main findings were as follows.

1. The values of the ET estimate of *MH D-DIF* were very close to the values obtained by averaging over 66 replications. The standard error of the difference between the ET estimate and the average over replications can be estimated using the standard errors from each

replication and the standard errors of the ET estimate (Appendix B). The standard errors of the averages over 66 replications were about .08 for these items; the standard errors of the ET estimates were about .04. This yielded values of about .09 for the standard errors of the difference between these two estimates. Only five of the differences between the two sets of estimates were found to be greater than .09 in magnitude. This is consistent with what would be expected if the differences were normally distributed, with a mean of zero. It is interesting that the ET approach yielded a more precise estimate of *MH D-DIF* than the average over 66 replications. In fact, for the Condition 5 pretest items, about 316 replications would have been required to match the precision of the ET estimate.

2. The distribution of *MH D-DIF* across items was examined. Based on the ET estimates, the mean and standard deviation were .17 and 1.27, respectively. Based on the mean *MH D-DIF* across replications, the across-item mean and standard deviation were found to be .12 and 1.32. The correlation across items between the two estimates of *MH D-DIF* was .997.

3. Estimates of the standard error of *MH D-DIF* were also considered. Here, three estimates were available--the ET estimate SE_{ET}*(MH D-DIF)*, the average of *SE(MH D-DIF)* over replications, and the observed standard deviation of *MH D-DIF* across replications. Differences among the estimates were very small. The average *SE(MH D-DIF)* tended to be slightly larger than the standard deviation of *MH D-DIF*, as found by Donoghue, Holland, and Thayer (1993). SE_{ET}*(MH D-DIF)* tended to be slightly smaller than the standard deviation. The across-item correlation between SE_{ET}*(MH D-DIF)* and the average *SE(MH D-DIF)* was

.992. Each of these estimates had correlations of about .8 with the standard deviation of *MH D-DIF*.

4. The estimated proportion of A, B, and C categorizations were examined. As shown in Table C-1, agreement between the two estimation methods on the estimated proportion of times the item would be labeled a "C," which was our main focus, were satisfactory for most items. An exception is item 6, which also had the largest discrepancy between methods in the estimated *MH D-DIF* statistics.

In summary, the ET method appears to give similar results to those obtained using more conventional estimation methods. In our study of the 15 pretest items, the ET estimates were much more precise than those that would have been obtained using an affordable number of replications.

Table C-1
Comparison of ET Estimates of DIF Statistics to Estimates Based on 66 Replications

Item		*MH D-DIF*	*SE (MH DIF)*		Estimated Percents in ETS DIF Categories A	B	C
1	ET	-2.12	.61		12.1	34.8	53.0
	Reps.	-2.12	.65[a]	(.65)	6.7	36.2	57.1
2	ET	-1.06	.62		71.2	22.7	6.1
	Reps.	-1.09	.65	(.58)	60.1	33.9	6.1
3	ET	0.08	.65		97.0	3.0	0.0
	Reps	-0.16	.68	(.68)	94.8	5.0	0.2
4	ET	1.25	.71		71.2	19.7	9.1
	Reps.	1.28	.75	(.78)	57.7	32.5	9.8
5	ET	2.55	.79		10.6	19.7	69.7
	Reps.	2.66	.85	(.79)	10.5	27.5	62.0
6	ET	-1.46	.63		30.3	37.9	31.8
	Reps.	-1.72	.67	(.71)	36.1	45.9	17.9
7	ET	-0.74	.61		77.3	21.2	1.5
	Reps	-0.78	.64	(.61)	77.2	20.9	1.9
8	ET	0.02	.60		100.0	0.0	0.0
	Reps	0.01	.62	(.59)	95.0	4.9	0.1
9	ET	1.01	.60		71.2	21.2	7.6
	Reps.	1.00	.62	(.63)	60.6	34.3	5.2
10	ET	1.98	.61		12.1	42.4	45.5
	Reps.	2.02	.64	(.66)	9.6	41.6	48.7
11	ET	-0.47	.70		92.4	7.6	0.0
	Reps.	-0.46	.73	(.78)	89.7	9.5	0.8
12	ET	-0.20	.68		93.9	6.1	0.0
	Reps.	-0.43	.73	(.76)	94.0	5.8	0.3
13	ET	0.10	.67		95.5	4.5	0.0
	Reps.	0.13	.70	(.65)	94.7	5.1	0.2
14	ET	0.54	.65		97.0	3.0	0.0
	Reps.	0.39	.68	(.62)	86.8	12.3	0.9
15	ET	1.11	.63		69.7	28.8	1.5
	Reps.	1.06	.66	(.58)	57.9	35.1	7.0

[a]The lefthand entry in the "Reps." row of this column is the average of the 66 values of *SE(MH D-DIF)* from the replications. The parenthesized value is the standard deviation of the 66 values of *MH D-DIF* from the replications.

Appendix D

Expected Proportions of A, B, and C DIF Results

Based on ETS Classification Rules

Suppose that the *MH D-DIF* statistic, *M*, has a normal distribution with mean μ and variance σ^2. Then estimates of the proportion of A, B, and C items (*PROPA*, *PROPB* and *PROPC*) can be derived as follows, using the *MH D-DIF* and *SE*(*MH D-DIF*) for all available data as estimates of μ and σ.

1. First, estimate *PROPBC*, the proportion of times the item will be a B or C item:

$$P(MH\,chi\text{-}square > 3.84) \equiv P\left((M/\sigma)^2 > 3.84\right)$$
$$= P(M/\sigma > 1.96) + P(M/\sigma < -1.96)$$
$$= P\left((M - \mu)/\sigma > (1.96 - \mu/\sigma)\right) + P\left((M - \mu)/\sigma < (-1.96 - \mu/\sigma)\right)$$
$$= P\left(Z > (1.96 - \mu/\sigma) + P\left(Z < (-1.96 - \mu/\sigma)\right)\right)$$

where Z is a standard normal variable.

$$P(|M| > 1) = P(M > 1) + P(M < -1)$$
$$= P\left((M - \mu)/\sigma > (1 - \mu)/\sigma\right) + P\left((M - \mu)/\sigma < (-1 - \mu)/\sigma\right)$$
$$= P\left(Z > (1 - \mu)/\sigma\right) + P\left(Z < (-1 - \mu)/\sigma\right)$$

$$K1BC \equiv \min\left(-1.96 - \mu/\sigma, (-1 - \mu)/\sigma\right)$$

$$K2BC \equiv \max\left(1.96 - \mu/\sigma, (1 - \mu)/\sigma\right)$$

$$PROPBC = P(Z > K2BC) + P(Z < K1BC).$$

2. Next, estimate *PROPC*, the proportion of times the item will be a *C* item:

$$P(|M| > 1.65\sigma + 1) = P(M > 1.65\sigma + 1) + P(M < -1.65\sigma - 1)$$
$$= P\left((M - \mu)/\sigma > 1.65 + (1 - \mu)/\sigma\right) + P\left((M - \mu)/\sigma < -1.65 - (1 + \mu)/\sigma\right)$$
$$= P\left(Z > 1.65 + (1 - \mu)/\sigma\right) + P\left(Z < -1.65 - (1 + \mu)/\sigma\right)$$
$$P(|M| > 1.5) = P(M > 1.5) + P(M < -1.5)$$
$$= P\left((M - \mu)/\sigma > (1.5 - \mu)/\sigma\right) + P\left((M - \mu)/\sigma < (-1.5 - \mu)/\sigma\right)$$
$$= P\left(Z > (1.5 - \mu)/\sigma\right) + P\left(Z < (-1.5 - \mu)/\sigma\right)$$

$$K1C \equiv \min\left(-1.65 - (1 + \mu)/\sigma, (-1.5 - \mu)/\sigma\right)$$

$$K2C \equiv \max\left(1.65 + (1 - \mu)/\sigma, (1.5 - \mu)/\sigma\right)$$

$$PROPC = P(Z > K2C) + P(Z < K1C)$$

3. Now calculate *PROPB* and *PROPA* by subtraction:

$$PROPB = PROPBC - PROPC$$

$$PROPA = 1 - PROPBC$$

Tables

Table 1

The 18 Administration Conditions

Condition	Focal Population	Sample size per item		Pool[b]
		Focal	Reference	
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

[b]Pool 1: no DIF, Pool 2: DIF uncorrelated with item difficulty, Pool 3: DIF positively correlated with item difficulty.

Table 2

Means and Standard Deviations of *ln a*, *b*, and *d* for Item Pools 1, 2, and 3

(Assuming a Multivariate Normal Distribution)

	Pool	
	1	2 and 3
ln a[a]		
mean	-.15	-.15
s.d.	.30	.30
b		
mean	0	0
s.d.	1.5	1.5
d		
mean	0	0
s.d.	0	.30

[a] *ln a* normal with mean -.15 and s.d. .30 corresponds to *a* log-normal with mean .9 and s.d. .28.

Table 3

Correlation Matrices of *ln a*, *b*, and *d* for Item Pools 1, 2, and 3

Pools 1 and 2

	ln a	*b*	*d*
ln a	1	.40	0
b		1	0
d			1

Pool 3

	ln a	*b*	*d*
ln a	1	.40	0
b		1	.40
d			1

Table 4

Correlations for True DIF (*ad*) and *MH D-DIF* Statistics Based on Three Types of Matching Variables [a]

Variables		Type of Correlation	Condition: / Pool: / Focal Group:	4 / 2 / N(-1,1)	6 / 3 / N(-1,1)	10 / 2 / N(0,1)	12 / 3 / N(0,1)	16 / 2 / N(.5,1)	18 / 3 / N(.5,1)	Median
θ̂-CAT	θ̂-75	Uncorrected		.83	.88	.89	.88	.91	.89	.89
		Corrected		.93	1.00[b]	1.00	.97	1.00[b]	.99	.99
θ̂-CAT	NR	Uncorrected		.85	.87	.89	.86	.90	.90	.88
		Corrected		.96	.99	.99	.96	1.00	1.00[b]	.99
θ̂-CAT	*ad*	Uncorrected		.96	.95	.98	.96	.99	.96	.96
		Corrected		.97	.96	.99	.97	1.00	.97	.97
θ̂-75	NR	Uncorrected		.99	.99	.99	.99	.99	.99	.99
		Corrected		1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]
θ̂-75	*ad*	Uncorrected		.84	.86	.88	.85	.90	.88	.87
		Corrected		.93	.97	.98	.93	.99	.98	.97
NR	*ad*	Uncorrected		.86	.87	.88	.84	.89	.89	.88
		Corrected		.95	.98	.98	.92	.99	.98	.98

[a]In this table, θ̂-CAT, θ̂-75, and NR refer to the *MH D-DIF* statistics that result from matching on expected true score based on the CAT, expected true score based on 75 item responses, and number-right score based on 75 items, respectively. Correlations are based on 71 items because 4 items were never administered in the CAT.

[b]Corrected value was greater than unity.

Table 5

Correlations for True DIF (*ad*) and *STD P-DIF* Statistics Based on Three Types of Matching Variables[a]

Variables		Type of Correlation	Condition: / Pool: / Focal Group:	4 / 2 / N(-1,1)	6 / 3 / N(-1,1)	10 / 2 / N(0,1)	12 / 3 / N(0,1)	16 / 2 / N(.5,1)	18 / 3 / N(.5,1)	Median
θ̂-CAT	θ̂-75	Uncorrected		.80	.80	.87	.88	.91	.86	.86
		Corrected		.89	.91	.96	.97	1.00[b]	.95	.95
θ̂-CAT	NR	Uncorrected		.81	.80	.88	.87	.91	.87	.87
		Corrected		.93	.94	.99	.96	1.00	.95	.95
θ̂-CAT	*ad*	Uncorrected		.96	.93	.98	.96	.99	.96	.96
		Corrected		.97	.94	.99	.96	.99	.96	.97
θ̂-75	NR	Uncorrected		.95	.95	.98	.98	.98	.99	.98
		Corrected		1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]
θ̂-75	*ad*	Uncorrected		.82	.79	.87	.87	.91	.88	.87
		Corrected		.90	.88	.95	.96	1.00	.97	.95
NR	*ad*	Uncorrected		.83	.81	.88	.87	.90	.88	.87
		Corrected		.94	.93	.98	.95	.99	.96	.95

[a]In this table, θ̂-CAT, θ̂-75, and NR refer to the *STD P-DIF* statistics that result from matching on expected true score based on the CAT, expected true score based on 75 item responses, and number-right score based on 75 items, respectively. Correlations are based on 71 items because 4 items were never administered in the CAT.

[b]Corrected value was greater than unity.

Table 6

Means and Standard Deviations of DIF Statistics for Three Types of Matching Variables[ab]

Matching Variable	Number of Items[c]	Condition: / Pool: / Focal Group:	4 / 2 / N(-1,1)	6 / 3 / N(-1,1)	10 / 2 / N(0,1)	12 / 3 / N(0,1)	16 / 2 / N(.5,1)	18 / 3 / N(.5,1)	Median
			MH D-DIF						
$\hat{\theta}$-CAT	71	mean	.00	.02	.02	.03	.01	.05	.02
		s.d.	.96	.89	.99	.94	1.02	.96	.96
$\hat{\theta}$-75	71	mean	-.02	.01	-.01	.00	-.01	-.04	-.01
		s.d.	.97	.90	.92	.96	1.02	.97	.97
NR	71	mean	-.02	-.02	-.02	-.02	-.02	-.08	-.02
		s.d.	.97	.88	.93	.99	1.03	.97	.97
NR	75	mean	.01	.01	.01	.00	.02	-.04	.01
		s.d.	.96	.87	.93	.98	1.02	.97	.97
			STD P-DIF × 10						
$\hat{\theta}$-CAT	71	mean	.01	.02	.01	.02	.00	.02	.01
		s.d.	.75	.72	.79	.76	.79	.77	.77
$\hat{\theta}$-75	71	mean	-.09	-.05	.03	.02	.03	.04	.02
		s.d.	.66	.62	.62	.64	.62	.65	.63
NR	71	mean	-.07	-.05	.00	-.01	.02	.01	.00
		s.d.	.71	.67	.64	.65	.62	.65	.64
NR	75	mean	-.04	-.02	.03	.02	.05	.04	.02
		s.d.	.70	.66	.65	.65	.63	.65	.65

[a]$\hat{\theta}$-CAT, $\hat{\theta}$-75, and NR refer to the DIF methods that match on expected true score based on the CAT, expected true score based on 75 item responses, and number right score based on 75 items, respectively. Correlations are based on 71 items because 4 items were never administered in the CAT.

[b]For conditions 4, 10, and 16, the mean value of *ad* across 71 items is -.004 and the standard deviation is .293. For conditions 6, 12, and 18, the mean and standard deviation are -.001 and .293, respectively.

[c]This column gives the number of items on which the tabled means and standard deviations are based.

Table 7

Pool 1:
Average *MH D-DIF* for Each Value of *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad*[b] 0
-1.95	0.0 (0.09) 21
-1.30	0.0 (0.07) 30
-.65	0.0 (0.06) 39
0	0.0 (0.11) 39
.65	0.0 (0.06) 33
1.30	0.0 (0.09) 33
1.95	0.0 (0.10) 18
Average	0.0 (0.08) 213

[a]The first entry in each cell is the average *MH D-DIF* for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, and the third entry is the number of item results over which the averages were computed.

[b]*ad* = 0 for all items in Pool 1.

Table 8

Pool 2:

Average *MH D-DIF* for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95			-1.3 (0.07) 3	-0.9 (0.08) 6	0.0 (0.10) 9	1.0 (0.10) 3				-0.3 (0.09) 21
-1.30			-1.3 (0.05) 3	-0.9 (0.08) 6	0.0 (0.06) 12	0.9 (0.09) 6	1.3 (0.06) 3			0.0 (0.07) 30
-.65			-1.3 (0.04) 6	-0.9 (0.08) 6	0.0 (0.05) 15	0.8 (0.10) 6	1.2 (0.04) 3		2.4 (0.05) 3	0.0 (0.06) 39
0	-2.1 (0.04) 3	-2.2 (0.30)[b] 3	-1.2 (0.04) 6	0.1 (0.58)[c] 3	0.0 (0.04) 9	0.9 (0.05) 6	1.2 (0.04) 3	1.8 (0.08) 3	2.5 (0.05) 3	0.1 (0.11) 39
.65			-1.2 (0.06) 6	-0.9 (0.06) 3	0.0 (0.08) 15	1.0 (0.05) 3	1.2 (0.04) 6			0.0 (0.07) 33
1.30		-1.8 (0.10) 3	-0.9 (0.06) 6	-0.5 (0.16) 3	0.0 (0.10) 12	1.0 (0.11) 3	1.1 (0.06) 6			-0.1 (0.09) 33
1.95			-0.7 (0.09) 3		0.0 (0.09) 6	1.1 (0.20) 3	0.8 (0.06) 6			0.3 (0.10) 18
Average	-2.1 (0.04) 3	-2.0 (0.20) 6	-1.1 (0.05) 33	-0.7 (0.14) 27	0.0 (0.07) 78	0.9 (0.09) 30	1.1 (0.05) 27	1.8 (0.08) 3	2.5 (0.05) 6	0.0 (0.08) 213

[a]The first entry in each cell is the average *MH D-DIF* for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, and the third entry is the number of item results over which the averages were computed.

[b]The average standard error is large because of the sparsity of data for Item 19.

[c]The average standard error is large because of the sparsity of data for Item 20.

Table 9

Pool 3:

Average *MH D-DIF* for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95		-1.6	-0.9	-0.7	0.3	1.3				-0.3
		(0.08)	(0.07)	(0.09)	(0.09)	(0.09)				(0.09)
		3	3	6	6	3				21
-1.30		-1.4	-0.9	-0.6	0.3	1.2	1.6			0.0
		(0.07)	(0.05)	(0.07)	(0.06)	(0.08)	(0.06)			(0.06)
		3	3	6	12	3	3			30
-.65	-2.2		-1.0	-0.7	0.2	1.0	1.5			-0.1
	(0.04)		(0.04)	(0.08)	(0.05)	(0.09)	(0.04)			(0.06)
	3		6	6	15	6	3			39
0			-1.1	-0.6	0.1	1.0	1.3			0.1
			(0.04)	(0.46)[b]	(0.04)	(0.06)	(0.04)			(0.11)
			6	6	15	6	6			39
.65			-1.4	-1.0	-0.1	0.8	1.1			-0.1
			(0.06)	(0.06)	(0.08)	(0.05)	(0.05)			(0.07)
			6	3	15	3	6			33
1.30			-1.1	-1.3	-0.3	0.6	0.9	1.6	2.1	0.2
			(0.05)	(0.09)	(0.11)	(0.08)	(0.05)	(0.11)	(0.08)	(0.09)
			3	3	12	3	6	3	3	33
1.95			-1.1		-0.5	0.6	0.6		1.5	0.3
			(0.08)		(0.11)	(0.17)	(0.06)		(0.06)	(0.09)
			3		3	3	6		3	18
Average	-2.2	-1.5	-1.1	-0.7	0.0	0.9	1.1	1.6	1.8	0.0
	(0.04)	(0.08)	(0.05)	(0.15)	(0.07)	(0.09)	(0.05)	(0.11)	(0.07)	(0.08)
	3	6	30	30	78	27	30	3	6	213

[a]The first entry in each cell is the average *MH D-DIF* for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, and the third entry is the number of item results over which the averages were computed.

[b]The average standard error is large because of the sparsity of data for Item 20.

Table 10

Pool 1:
Average *STD P-DIF* × 10 for Each Value of *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad*[b] 0
-1.95	0.00 (0.07) 21
-1.30	-0.01 (0.06) 30
-.65	0.00 (0.05) 39
0	-0.01 (0.08) 39
.65	0.01 (0.06) 33
1.30	0.01 (0.08) 33
1.95	0.02 (0.09) 18
Average	0.00 (0.07) 213

[a]The first entry in each cell is the average *STD P-DIF*, multiplied by 10, for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, multiplied by 10, and the third entry is the number of item results over which the averages were computed.

[b]*ad* = 0 for all items in Pool 1.

Table 11

Pool 2:

Average *STD P-DIF* × 10 for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty	Value of *ad*									
(*b*)	-.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95			-0.70	-0.74	0.04	0.72				-0.19
			(0.04)	(0.07)	(0.08)	(0.07)				(0.07)
			3	6	9	3				21
-1.30			-0.92	-0.83	0.02	0.81	0.84			0.00
			(0.04)	(0.07)	(0.05)	(0.08)	(0.04)			(0.06)
			3	6	12	6	3			30
-.65			-0.98	-0.78	-0.04	0.64	0.89		1.70	0.01
			(0.03)	(0.07)	(0.05)	(0.08)	(0.04)		(0.03)	(0.05)
			6	6	15	6	3		3	39
0	-1.83	-1.27	-0.91	0.08	-0.03	0.76	0.81	1.49	2.02	0.07
	(0.04)	(0.20)[b]	(0.03)	(0.36)[c]	(0.04)	(0.04)	(0.03)	(0.06)	(0.04)	(0.08)
	3	3	6	3	9	6	3	3	3	39
.65			-1.00	-0.75	-0.01	0.80	0.97			0.00
			(0.05)	(0.05)	(0.07)	(0.04)	(0.04)			(0.06)
			6	3	15	3	6			33
1.30		-1.66	-0.72	-0.41	-0.02	0.82	0.90			-0.09
		(0.09)	(0.05)	(0.14)	(0.09)	(0.10)	(0.06)			(0.08)
		3	6	3	12	3	6			33
1.95			-0.58		0.00	0.97	0.63			0.28
			(0.07)		(0.08)	(0.18)	(0.05)			(0.08)
			3		6	3	6			18
Average	-1.83	-1.46	-0.85	-0.64	-0.01	0.77	0.84	1.49	1.86	0.01
	(0.04)	(0.15)	(0.04)	(0.11)	(0.06)	(0.08)	(0.04)	(0.06)	(0.04)	(0.07)
	3	6	33	27	78	30	27	3	6	213

[a]The first entry in each cell is the average *STD P-DIF*, multiplied by 10, for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, multiplied by 10, and the third entry is the number of item results over which the averages were computed.

[b]The average standard error is large because of the sparsity of data for Item 19.

[c]The average standard error is large because of the sparsity of data for Item 20.

Table 12

Pool 3:

Average *STD P-DIF* × 10 for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty	Value of *ad*									
(*b*)	-.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95		-1.35	-0.49	-0.55	0.21	0.98				-0.22
		(0.07)	(0.04)	(0.08)	(0.06)	(0.07)				(0.07)
		3	3	6	6	3				21
-1.30		-1.28	-0.66	-0.54	0.22	1.09	1.04			0.00
		(0.07)	(0.04)	(0.06)	(0.05)	(0.08)	(0.04)			(0.06)
		3	3	6	12	3	3			30
-.65	-1.74		-0.78	-0.62	0.18	0.85	1.12			-0.06
	(0.03)		(0.04)	(0.07)	(0.05)	(0.08)	(0.04)			(0.05)
	3		6	6	15	6	3			39
0			-0.87	-0.38	0.07	0.84	1.05			0.13
			(0.03)	(0.32)[b]	(0.04)	(0.06)	(0.04)			(0.08)
			6	6	15	6	6			39
.65			-1.17	-0.85	-0.09	0.65	0.91			-0.11
			(0.05)	(0.05)	(0.07)	(0.04)	(0.04)			(0.06)
			6	3	15	3	6			33
1.30			-0.88	-1.13	-0.22	0.49	0.71	1.25	1.60	0.17
			(0.04)	(0.09)	(0.09)	(0.08)	(0.04)	(0.09)	(0.07)	(0.07)
			3	3	12	3	6	3	3	33
1.95			-0.87		-0.46	0.60	0.45		1.20	0.23
			(0.07)		(0.10)	(0.16)	(0.05)		(0.05)	(0.08)
			3		3	3	6		3	18
Average	-1.74	-1.31	-0.85	-0.62	0.03	0.80	0.84	1.25	1.40	0.02
	(0.03)	(0.07)	(0.04)	(0.12)	(0.06)	(0.08)	(0.04)	(0.09)	(0.06)	(0.07)
	3	6	30	30	78	27	30	3	6	213

[a]The first entry in each cell is the average *STD P-DIF*, multiplied by 10, for the indicated values of *ad* and *b*, the second entry is the average standard error of the estimate, multiplied by 10, and the third entry is the number of item results over which the averages were computed.

[b]The average standard error is large because of the sparsity of data for Item 20.

Table 13

Pool 1:
Average Expected Percent of C Results for Each Value of b[a]

Item Difficulty (b)	Value of ad[b] 0
-1.95	0.0
	0.2
	21
-1.30	0.0
	0.1
	30
-.65	0.0
	0.1
	39
0	0.0
	0.1
	39
.65	0.0
	0.1
	33
1.30	0.0
	0.1
	33
1.95	0.0
	0.1
	18
Average	0.0
	0.1
	213

[a]The first entry in each cell is the average expected percent of C results for the indicated values of ad and b in the 900, 100 sample size condition, the second entry is the average percent for the 500, 500 sample size condition, and the third entry is the number of item results over which the averages were computed.

[b]$ad = 0$ for all items in Pool 1.

Table 14

Pool 2:
Average Expected Percent of C Results for Each Combination of *ad* and *b*[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95			10.3	3.6	0.2	4.2				3.2
			15.1	3.3	0.0	3.9				3.6
			3	6	9	3				21
-1.30			11.9	3.4	0.1	3.4	11.1			3.7
			19.5	2.9	0.0	2.7	18.7			4.9
			3	6	12	6	3			30
-.65			11.4	3.5	0.1	2.2	9.8		67.2	8.6
			18.7	3.2	0.0	1.4	15.5		97.3	12.3
			6	6	15	6	3		3	39
0	60.5	49.6	8.2	0.5	0.1	3.7	7.8	38.3	78.3	19.9
	93.9	83.0	11.6	0.1	0.0	3.1	11.6	75.3	99.5	30.2
	3	3	6	3	9	6	3	3	3	39
.65			8.5	2.7	0.1	4.3	8.7			3.8
			12.0	1.8	0.0	4.0	12.8			5.0
			6	3	15	3	6			33
1.30		44.8	4.1	0.8	0.1	5.2	8.1			6.9
		81.4	4.0	0.2	0.0	5.7	11.5			10.8
		3	6	3	12	3	6			33
1.95			2.0		0.1	7.5	3.3			2.7
			1.2		0.0	11.3	2.8			3.0
			3		6	3	6			18
Average	60.5	47.2	8.1	2.8	0.1	4.0	7.7	38.3	72.7	7.9
	93.9	82.2	11.7	2.3	0.0	3.9	11.1	75.3	98.4	11.5
	3	6	33	27	78	30	27	3	6	213

[a]The first entry in each cell is the average expected percent of C results for the indicated values of *ad* and *b* in the 900, 100 sample size condition, the second entry is the average percent for the 500, 500 sample size condition, and the third entry is the number of item results over which the averages were computed.

Table 15

Pool 3:
Average Expected Percent of C Results for Each Combination of *ad* and *b*[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.52	-.35	-.26	0	.26	.35	.52	.70	Average
-1.95		26.8	3.2	1.2	0.4	12.5				6.6
		52.1	2.5	0.5	0.1	21.9				11.1
		3	3	6	6	3				21
-1.30		17.6	3.6	0.9	0.3	10.0	22.0			5.6
		34.1	3.1	0.2	0.0	15.3	45.0			9.8
		3	3	6	12	3	3			30
-.65	63.8		4.2	1.7	0.1	5.5	18.8			8.2
	95.2		4.0	0.9	0.0	6.2	37.8			11.9
	3		6	6	15	6	3			39
0			6.4	2.6	0.1	5.0	14.2			4.4
			8.0	2.6	0.0	5.3	26.0			6.4
			6	6	15	6	6			39
.65			15.6	4.6	0.1	2.3	7.0			4.8
			27.7	4.4	0.0	1.2	9.1			7.2
			6	3	15	3	6			33
1.30			8.6	12.1	0.3	0.8	3.6	26.3	54.7	10.1
			12.0	19.4	0.1	0.2	3.2	55.1	92.4	16.9
			3	3	12	3	6	3	3	33
1.95			8.4		0.7	1.2	1.0		27.1	6.6
			11.8		0.1	0.4	0.3		48.1	10.2
			3		3	3	6		3	18
Average	63.8	22.2	7.6	3.0	0.2	5.3	9.3	26.3	40.9	6.6
	95.2	43.1	10.9	3.2	0.0	6.9	16.0	55.1	70.3	10.4
	3	6	30	30	78	27	30	3	6	213

[a]The first entry in each cell is the average expected percent of C results for the indicated values of *ad* and *b* in the 900, 100 sample size condition, the second entry is the average percent for the 500, 500 sample size condition, and the third entry is the number of item results over which the averages were computed.

Table 16

Pretest Items (Pool 1):

Average *MH D-DIF* for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.35	0	.35	.70	Average
-1.30	-2.4	-1.3	0.0	1.2	2.5	0.0
	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
0	-1.9	-1.0	0.0	1.1	2.2	0.1
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
1.30	-1.0	-0.6	0.0	0.7	1.6	0.1
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Average	-1.8	-0.9	0.0	1.0	2.1	0.1
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)

[a]The first entry in each cell is the average *MH D-DIF* for the indicated values of *ad* and *b* and the second entry is the average standard error of the estimate. Each cell average is based on 3 item results.

Table 17

Pretest Items (Pool 1):

Average *STD P-DIF* × 10 for Each Combination of *ad* and *b* in the 500, 500 Sample Size Condition[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.35	0	.35	.70	Average
-1.30	-1.33	-0.65	-0.03	0.47	0.90	-0.13
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
0	-1.41	-0.71	-0.01	0.75	1.47	0.02
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
1.30	-0.73	-0.40	0.05	0.54	1.20	0.13
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Average	-1.16	-0.59	0.00	0.58	1.20	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)

[a]The first entry in each cell is the average *STD P-DIF*, multiplied by 10, for the indicated values of *ad* and *b* and the second entry is the average standard error of the estimate, multiplied by 10. Each cell average is based on 3 item results.

Table 18

Pretest Items (Pool 1):

Average Expected Percent of C Results for Each Combination of *ad* and *b*[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.35	0	.35	.70	Average
-1.30	65.5	10.0	0.4	7.4	41.9	25.0
	93.5	15.5	0.0	10.5	82.0	40.3
0	41.4	4.3	0.1	6.3	53.8	21.2
	79.6	4.6	0.0	8.1	88.9	36.2
1.30	8.0	1.1	0.1	1.9	28.3	7.9
	9.8	0.3	0.0	0.9	49.2	12.1
Average	38.3	5.1	0.2	5.2	41.4	18.0
	61.0	6.8	0.0	6.5	73.4	29.5

[a]The first entry in each cell is the average expected percent of C results for the indicated values of *ad* and *b* in the 900, 100 sample size condition and the second entry is the average percent for the 500, 500 sample size condition. Each cell average is based on 3 item results.

Table 19

Pretest Items (Pool 2):

Average Expected Percent of C Results for Each Combination of *ad* and *b*[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.35	0	.35	.70	Average
-1.30	65.3	9.6	0.4	7.4	41.8	24.9
	93.7	14.9	0.0	10.3	82.2	40.2
0	43.4	5.1	0.1	6.5	53.4	21.7
	81.4	6.3	0.0	8.5	89.2	37.1
1.30	7.7	1.0	0.1	1.8	27.6	7.6
	9.6	0.3	0.0	0.9	47.2	11.6
Average	38.8	5.2	0.2	5.2	40.9	18.1
	61.6	7.1	0.0	6.5	72.8	29.6

[a]The first entry in each cell is the average expected percent of C results for the indicated values of *ad* and *b* in the 900, 100 sample size condition and the second entry is the average percent for the 500, 500 sample size condition. Each cell average is based on 3 item results.

Table 20

Pretest Items (Pool 3):

Average Expected Percent of C Results for Each Combination of *ad* and *b*[a]

Item Difficulty (*b*)	Value of *ad* -.70	-.35	0	.35	.70	Average
-1.30	53.6	5.2	0.6	11.5	52.0	24.6
	87.2	6.5	0.1	19.4	91.2	40.9
0	40.4	3.4	0.1	9.0	61.5	22.9
	78.8	3.2	0.0	14.3	94.2	38.1
1.30	10.5	1.5	0.1	1.4	24.9	7.7
	13.4	0.5	0.0	0.6	45.9	12.1
Average	34.8	3.3	0.3	7.3	46.1	18.4
	59.8	3.4	0.0	11.5	77.1	30.3

[a]The first entry in each cell is the average expected percent of C results for the indicated values of *ad* and *b* in the 900, 100 sample size condition and the second entry is the average percent for the 500, 500 sample size condition. Each cell average is based on 3 item results.

Table 21

Median and Interquartile Range of Examinee Residuals $\left(\hat{\theta}_{CAT} - \theta\right)$ for Samples of 1,000[a]

Group	Pool 1	Pool 2	Pool 3
Reference	-0.032		
	0.450		
Focal: N(-1,1)	-0.038	-0.032	-0.099
	0.515	0.523	0.569
Focal: N(0,1)	-0.031	-0.014	-0.044
	0.458	0.487	0.502
Focal: N(0.5,1)	-0.036	-0.048	-0.011
	0.445	0.472	0.488

[a] Standard errors of medians are approximately 0.02.

Table 22

Median and Interquartile Range of Examinee Residuals $(\hat{\theta}_{75} - \theta)$ for Samples of 1,000[a]

Group	Pool 1	Pool 2	Pool 3
Reference	-0.030		
	0.379		
Focal: N(-1,1)	-0.006	-0.003	-0.063
	0.421	0.425	0.413
Focal: N(0,1)	-0.025	-0.009	-0.037
	0.333	0.337	0.391
Focal: N(0.5,1)	-0.028	-0.026	-0.013
	0.382	0.331	0.360

[a] Standard errors of medians are approximately 0.01.