DOCUMENT RESUME

ED 351 384 TM 019 216

AUTHOR	Nandakumar, Ratna
TITLE	Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF.
INSTITUTION	Illinois Univ., Urbana. Dept. of Statistics.
SPONS AGENCY	Office of Naval Research, Arlington, Va.
REPORT NO	1992-4; ONR-4421-548
PUB DATE	15 Aug 92
CONTRACT	N00014-90-J-1940
NOTE	39p.; Paper to be published in the "Journal of Educational Measurement."
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Black Students; Comparative Testing; Computer Simulation; Computer Software; Equations (Mathematics); Females; *Item Bias; Males; *Mathematical Models; Racial Differences; *Research Methodology; Sex Differences; Test Interpretation; Test Items; *Test Results; White Students
IDENTIFIERS	American College Testing Program; *Amplification (Simulatenous Differen Item Func); *Cancellation (Simultaneous Differen Item Func); National Assessment of Educational Progress; SIBTEST (Computer Program)

ABSTRACT

The phenomenon of simultaneous differential item functioning (DIF) amplification and cancellation and the role of the SIBTEST computer program in detecting it were studied. A variety of simulated test data was generated for this purpose. In addition, the following real test data were used: (1) American College Testing program data for 2,115 males and 2,885 females in mathematics; (2) National Assessment of Educational Progress (NAEP) history test data for 1,225 males and 1,215 females; and (3) NAEP data for 1,711 whites and 447 blacks. The results from both simulated and real data, as the theory of R. Shealy and W. F. Stout suggests, show that SIBTEST is effective in detecting DIF amplification and cancellation (partially or fully) at the test score level. Finally, methodological and substantive implications of DIF amplification and cancellation are discussed. Ten tables present analysis results. (SLD)

Simultaneous DIF Amplification and Cancellation: Sheady-Stout's Test for DIF

Ratna Nandakumar[1]
Department of Educational Studies
University of Delaware

August 15, 1992

Prepared for the Cognitive Science Research Program, Cognitive and Neural Sciences Division, Office of Naval Research, under grant number N00014-90-J-1940, 4421-548. Approved for public release, distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[1] The author would like to convey special thanks to William Stout for his insightful suggestions on this research and to Louis Roussos for programming assistance.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
15 August 1992

3. REPORT TYPE AND DATES COVERED
Technical: 1990-93

4. TITLE AND SUBTITLE
Simultaneous DIF Amplification and Cancellation:
Shealy-Stout's Test for DIF

5. FUNDING NUMBERS
N00014-90-J-1940,

6. AUTHOR(S)
Ratna Nandakumar

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Statistics
University of Illinois
725 South Wright Street
Champaign, IL 61820

8. PERFORMING ORGANIZATION REPORT NUMBER
1992 - No. 4

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Cognitive Sciences Program
Office of Naval Research
800 N. Quincy
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
4421-548

11. SUPPLEMENTARY NOTES
To be published in Journal of Educational Measurement

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
See reverse

14. SUBJECT TERMS
See reverse

15. NUMBER OF PAGES
33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
unclassified

20. LIMITATION OF ABSTRACT
UL

Simultaneous DIF Amplification and Cancellation: Shealy–Stout's Test for DIF

Abstract

The present study investigates the phenomena of simultaneous DIF amplification and cancellation and SIBTEST's role in detecting such. A variety of simulated test data were generated for this purpose. In addition, real test data from various sources were used. The results from both simulated as well as real test data, as Shealy and Stout's theory suggests, show that the SIBTEST is effective in assessing the DIF amplification and cancellation (partially or fully) at the test score level. Finally, methodological and substantive implications of DIF amplification and cancellation are discussed.

Subject terms: SIBTEST, DIF, item bias, test bias, bias amplification, bias cancellation.

Simultaneous DIF Amplification and Cancellation: Shealy–Stout's Test for DIF

Studies of bias have been widely prevalent in educational measurement since the 1960s. Early attempts to study bias in tests were largely based on the notion of predictive validity. Consequently, a number of regression models were developed, based on different definitions of fairness, in order to achieve fair employment selection and college admissions (Peterson and Novick, 1976). Since the advent of item response theory (IRT), however, study of bias and differential item functioning (DIF) at the item leve! has gained much popularity. Several methodologies have been developed by various researchers to study item bias and DIF (for descriptions and/or comparisons of different procedures, see for example, Angoff, 1982; Cleary & Hilton, 1968; Dorans & Kulick, 1983, 1986; Hambleton & Rogers, 1989; Holland & Thayer, 1988; Hunter, 1975; Ironson, 1982; Lord, 1980; Raju, 1988; Reynolds, 1982; Scheuneman, 1979; Shealy & Stout, 1992b; Shepard, Camilli, & Averill, 1981; Swaminathan & Rogers, 1990; Wainer, Sireci, & Thissen, 1991).

These procedures can usually be used in an effort to detect either item bias or DIF. The subtle distinction between the closely related concepts of bias and DIF can be explained as follows. In the conceptualization of "item bias", it is generally assumed that the validity of some items of the test could be questionable while the rest of the items are considered valid. That is, these items of questionable validity could contribute to test score differences between groups of examinees with *equal ability*. In DIF analyses, however, it is conceptualized that some items could contribute to test score differences between two groups of examinees *matched* according to some criterion about which no validity claim is made. For example, examinees could be matched upon total test score with no accompanying claim of validity for the items of the test. Therefore, in item bias analyses, the construct validity of the matching subtest needs to be established while in DIF analyses it is not needed. In this sense item bias is a special case of DIF. Several biased items acting in concert produce test bias, and several DIF items acting in concert produce DTF (differential test functioning). Shealy and Stout (1992b) have further discussed the differences between bias and DIF analyses in a more detailed manner.

One of the recently developed IRT based methodologies for detecting item/test bias or DIF/DTF has been developed by Shealy and Stout (1992a,1992b). Known as SIBTEST (SIB denotes simultaneous item bias), it is a statistical test to simultaneously detect bias present in one or more items of a test. SIBTEST is an outgrowth of the multidimensional IRT modeling of test bias as presented in Shealy and Stout (1992a), and it is the first among IRT based procedures to allow the simultaneous testing for bias present in more than one item. The phenomenon of simultaneous item bias is said to occur when several biased items acting in concert affect the test score differentially for the different examinee subpopulations, resulting in test bias. In part, because of its multidimensional modeling approach, SIBTEST has several distinct features. First, single item bias as well as simultaneous item bias can be detected. Second, a formal distinction can be made between genuine test bias and impact, which is due to ability differences between groups in the ability intended to be measured (Ackerman, 1991a, Dorans, 1989). Third, the underlying psychological (cognitive) mechanisms that produce bias can be explicitly addressed through consideration of the *target ability* as contrasted with *nuisance determinants*. The target ability θ is the ability intended to be measured by the test, the nuisance determinant(s) η is an ability or construct not intended to be measured by the test but influencing the responses to one or more items.

One of the major advantages of considering simultaneous item bias is that it is possible to study item bias amplification and item bias cancellation. Bias amplification is illustrated by the following: if a set of individual items is each biased against males, then one can study the effect of the bias collectively against males at the overall test score level. Bias cancellation is illustrated by the following: if one set of individual items is each biased against males and another set of items is each biased against females, then it is possible that at the overall test score level the respective biases might cancel each other out. In any bias study one should investigate both of these possibilities. The phenomenon of item bias cancellation has been previously studied empirically by Drasgow (1987), Roznowski (1987), and Reith and Roznowski (1991).

Reith and Roznowski (1991) and Roznowski (1987) have studied the effect of biased items on the predictive validity of the test. They concluded that inclusion of biased items in the test can actually contribute to increased predictive validity when the sources of bias are diverse and multiply determined. They argue that, although items with non–trait (but trait–relevant) variance may manifest bias at the item level, nonetheless, several such items can actually improve the amount of variance explained by the trait at the test score level (here "trait" refers to the ability of interest). This is because, at the test score level, the amount of non–trait variance diminishes while the trait variance increases, thus improving the predictive validity. Thus, the removal of biased items might sometimes be considered to be detrimental to the predictive validity of the test.

Drasgow (1987) has shown, using Lord's chi–square item bias statistic, that several biased items of ACT mathematics usage and English usage tests, biased in different directions (some against Whites, some against Blacks, some against Hispanics, etc.), had no cumulative bias effect on the expected number–correct score. That is, there were no consistent differences in the test scores across groups. This was attributed to bias cancellation across groups. Humphreys (1970, 1986) has long recommended deliberate inclusion of diverse non–trait determinants in test items in order to diminish the biasing influence of any particular non–trait ability at the test score level. These studies clearly show that the study of the effect of amplification or cancellation of biased or DIF items at the test score level is a significant problem. Shealy and Stout (1992a) directly address these issues by modeling bias in a multidimensional frame work and considering the simultaneous influence of several biased items at once. According to them, the presence of multidimensionality is a prerequisite for bias. If test data can be modeled by a unidimensional or an essentially unidimensional (Stout, 1990) model, then bias cannot exist. The concept of bias in a multidimensional frame work has also been emphasized by Shepard (1982), Kok (1988) and others. As noted before, the SIBTEST procedure is an outgrowth of the multidimensional modeling of bias.

Shealy and Stout (1992a, 1992b) have demonstrated through simulation studies the ability of SIBTEST to detect unidirectional bias; that is, bias against the same group regardless of the level of target ability θ. In their simulations, they used two– and three–parameter logistic models with varying sample sizes and differing degrees of induced bias. The findings showed that SIBTEST displayed good adherence to the nominal level of significance in cases of no bias and good power in cases where one or more items were biased, even when the amount of bias was fairly small. In cases of single item bias studies, the performance of SIBTEST was compared to that of the Mantel–Haenszel statistic. Both the SIBTEST and the Mantel–Haenszel procedures produced consistent results with respect to the direction and the amount of estimated bias.

The purpose of this paper is to define the concepts of DIF amplification and DIF cancellation and to investigate the power of SIBTEST to address these phenomena. A series of real data and simulation data are used for this purpose. In case of single item analyses, SIBTEST results are compared with the Mantel–Haenszel results. Also, a brief description of the SIBTEST procedure is provided.

Description of SIBTEST Procedure

In this section, for ease of presentation, we will assume the bias viewpoint rather than the DIF/DTF viewpoint. It is vital, however, to realize that a similar presentation could have been given using the DIF/DTF perspective. As discussed before, the interpretations of SIBTEST results have either a test bias or a DTF interpretation, depending upon the level of user assumptions about the validity of the matching subtest items. In particular, SIBTEST can be used as a DIF procedure if desired.

Two groups (or subpopulations) of interest, the reference group (R) and the focal group (F), are assumed to take a given test. The complete latent space $\underline{\theta}$ underlying the test items is assumed to be multidimensional: $\{\underline{\theta} = (\theta, \underline{\eta})\}$, where θ is the target ability, intended to be measured by the test, and $\underline{\eta}$ is the nuisance ability vector (possibly multidimensional), not intended to be measured by test items. For example, in an English vocabulary test, it is possible that some items are male oriented, such as those requiring knowledge of sports, and some other items are female oriented, such as those requiring knowledge of domestics. In a situation like this, English vocabulary skill is the intended to be measured ability (θ). Knowledge of sports (η_1) and knowledge of domestics (η_2) are nuisance abilities. Let $\underline{U}$ denote the test response vector and h($\underline{U}$) the test scoring method. Number correct is used as the scoring method throughout this paper. It is assumed that all items of the given test measure the target ability θ, and some items (biased items) measure both target ability and one or more nuisance abilities $\underline{\eta}$. It is also assumed that the usual IRT assumptions of local independence, monotonicity, and group invariance hold with respect to $\underline{\theta}$ and that this collection of assumptions do not hold for any subset of components of $\underline{\theta}$.

The statistical procedure for testing the null hypothesis of no test bias is briefly explained below, for details see Shealy and Stout (1992b). The hypothesis can be stated as:

$$H_0: \beta_U = 0 \text{ vs. } H_1: \beta_U > 0,$$

where β_U is a parameter denoting the amount of unidirectional test bias against the focal group. Unidirectional bias occurs if the probability of answering an item(s) is consistently higher (lower) for one group compared to the other, over all levels of ability θ. That is, marginal item characteristic curves[1] for the two groups do not cross as θ varies over the ability range. Let $X = \Sigma_1^n U_i$ be the total score on the valid subtest, which by definition, consists of n items the user is willing to assume measure the target ability. Let $Y = \Sigma_{n+1}^{N} U_i$ be the total score on the studied subtest which consists of one or more items measuring target and possibly nuisance abilities. It is assumed that, for long tests, examinees with the same valid subtest score are of approximately equal target ability θ and thus are comparable. Following this logic, examinees within reference and focal groups are subgrouped according to their total score on the valid subtest. Examinees with the same valid subtest score are then compared across reference and focal groups on their performance on the studied subtest item(s). The test statistic, which is a sort of standardization index (see Dorans & Kulick, 1986), for testing the null hypothesis of no bias is then given by

$$B = \frac{\hat{\beta}_U}{\sigma(\hat{\beta}_U)}, \qquad (1)$$

where $\hat{\beta}_U = \sum_0^K \hat{p}_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$, and $\hat{p}_k$ is the proportion among focal group[2] examinees attaining $X=k$ on the valid subtest. $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ are the "adjusted" means of the studied subtest for examinees with a valid subtest score of $X=k$ ($k=0,1,\ldots,n$) in the reference and focal groups respectively. Because the procedure must work for short as well as long tests, these means are adjusted for differences in the θ distributions between reference and focal groups arising from short test lengths (for example, 25 items), and inherent differences in the θ distributions for the two groups (for details, see regression correction in Shealy &

Stout, 1992b). $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$ given by

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=0}^{n} \hat{p}_k^2 \left(\frac{1}{J_{Rk}}\hat{\sigma}^2(Y|\,k,R) + \frac{1}{J_{Fk}}\hat{\sigma}^2(Y|\,k,F)\right)\right)^{1/2},$$

where $\hat{\sigma}^2(Y|\,k,g)$ is the sample variance of the studied subtest for examinees in group g (R or F) with a total score of k on the valid subtest; and J_{Rk} and J_{Fk} are the sample sizes in the reference and focal groups respectively with a total score of k on the valid subtest.

The null hypothesis of no bias is rejected with error rate α if the value of B exceeds the upper 100(1−α)th percentile point of the standard normal distribution. $\hat{\beta}_U$ is also the statistic used tr ..timate the amount of unidirectional bias β_U. For example, a $\hat{\beta}_U$ value of 0.1 indicates that the average difference in the expected total test scores between reference and focal group examinees of similar ability is 0.1. If this is the result of a single studied item with the reminder of the items assumed valid, then $\hat{\beta}_U = 0.1$ is the estimated difference in the probability of getting the studied item correct between reference and focal group examinees of similar ability. Positive values of $\hat{\beta}_U$ indicate bias against the focal group and negative values of $\hat{\beta}_U$ indicate bias against the reference group. Simulation studies by Shealy and Stout (1992b) showed that B has good statistical properties such as good adherence to the nominal significance level and high po. er.

Simulation Study

Details about Simulations

In order to investigate amplification and cancellation of DIF and the use of SIBTEST to detect such, a simulation study was designed to model realistic situations. Item parameters (a_i, b_i, c_i) of valid subtests were obtained from the literature and the item parameters of studied subtests were hand selected to control the amount of DIF present. The estimated item parameters from the SAT–Verbal (Drasgow, 1987) were used for valid subtests. The parameters of the studied subtest items (that is, DIF items) are listed in Table 1. Item parameters of studied subtests were selected such that the difficulty parameters were all centered around zero, with varying discrimination parameters for θ, η_1 and η_2. All studied subtest items, except the last three, are influenced by θ and η_1. The last three items are influenced by θ and η_2. The guessing level is fixed to 0.2 for all items. For amplification studies, only items with nuisance ability η_1 were used. For the amplification and cancellation study, both, items with nuisance ability η_1, and items with nuisance ability η_2 were used.

Amplification Study

The target and the nuisance abilities were generated from a bivariate normal distribution as follows. For notational simplicity the subscript for η_1 is dropped.

$$\begin{pmatrix}\Theta\,|\,g\\ \eta\,|\,g\end{pmatrix} - N\left[\begin{pmatrix}\mu_{\theta g}\\ \mu_{\eta g}\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right], \tag{2}$$

where ρ is the correlation between θ and η for group g, which is set at 0.5 for both groups (different values of ρ across groups tends to produce bidirectional DIF). As can be seen the variances $\sigma^2(\theta\,|\,g)$ and $\sigma^2(\eta|\,g)$ were set at 1. The means $\mu_{\theta g}$ and $\mu_{\eta g}$ for each group were determined through specification of other parameters as follows.

Target ability difference between the reference and focal groups is denoted by

$$d_T = \frac{\mu_{\theta R} - \mu_{\theta F}}{\sigma_{\theta P}}, \tag{3}$$

where

$$\sigma^2_{\theta P} = \alpha_R \sigma^2(\Theta | R) + \alpha_F \sigma^2(\Theta | F);\ \alpha_R = \frac{J_R}{J_R + J_F} \text{ and } \alpha_F = \frac{J_F}{J_R + J_F};$$

and J_R and J_F denote sample sizes in reference and focal groups respectively. $\sigma^2_{\theta P}$ is the weighted average of the variances of reference and focal groups on the target ability. Since $\sigma^2(\Theta | R)$ and $\sigma^2(\Theta | F)$ were taken as 1 in simulation studies (see Equation 2), $d_T = \mu_{\theta R} - \mu_{\theta F}$. That is, d_T is a measure of how much the two groups differ in target ability distributions (same as impact).

Another criterion for choosing $\mu_{\theta R}$ and $\mu_{\theta F}$ was that the average difficulty level ($\bar{b}$) of the valid subtest items was assumed equal to the average target ability pooled across groups:

$$\bar{b} = E[\Theta] = \alpha_R \mu_{\theta R} + \alpha_F \mu_{\theta F}. \tag{4}$$

That is, on average the difficulty of the valid subtest items is assumed to be well matched with the pooled average target ability of the two groups. By specifying d_T and $\bar{b}$, Equations 3 and 4 together determine $\mu_{\theta R}$ and $\mu_{\theta F}$. Parameters $\mu_{\eta R}$ and $\mu_{\eta F}$ were determined as follows.

<u>Potential for DIF C_β</u> is defined as the difference between the conditional expectation of η for the two groups, given by

$$\begin{aligned} C_\beta &= E[\eta_R | \theta] - E[\eta_F | \theta] \\ &= (\mu_{\eta R} - \mu_{\eta F}) + (\rho \frac{\sigma_{\eta R}}{\sigma_{\theta R}})(\theta - \mu_{\theta R}) - (\rho \frac{\sigma_{\eta F}}{\sigma_{\theta F}})(\theta - \mu_{\theta F}) \end{aligned}$$

Following Equation 2 and 3

$$C_\beta = (\mu_{\eta R} - \mu_{\eta F}) - \rho d_T \tag{5}$$

Another criterion for choosing the means of η is that, for an "average" value of target ability (Θ=0) we assume the conditional nuisance ability to be centered around the chosen target ability value for the two groups. Namely,

$$E[\eta_R | \Theta=0] = -E[\eta_F | \Theta=0]$$

That is,

$$(\mu_{\eta R} - \rho\mu_{\theta R}) = -(\mu_{\eta F} - \rho\mu_{\theta F}) \qquad (6)$$

Once $\mu_{\theta R}$ and $\mu_{\theta F}$ are known, by specifying C_β, $\mu_{\eta R}$ and $\mu_{\eta F}$ can be determined from Equations 5 and 6.

The choice of values for C_β in the simulations were guided by the desired amount of the estimated DIF, $\hat{\beta}_U$. In other words, values of C_β were chosen so that the amount of estimated DIF would be "small" ($0 \le \hat{\beta}_U < 0.05$), "moderate" ($0.05 \le \hat{\beta}_U < 0.1$), or "large" ($\hat{\beta}_U \ge 0.1$). From the practical viewpoint, the standard used to determine what is meant by small, moderate, or large DIF was based on observed delta values of the Mantel–Haenszel statistic Δ_{MH} (Holland & Thayer, 1988). An approximate empirical relationship between Δ_{MH} and β_U is given by

$$\beta_U \simeq -\Delta_{MH}/10 \qquad (7)$$

Recall that β_U is a measure of the average difference in expected test scores between reference and focal group members of similar ability. That is, β_U as estimated by $\hat{\beta}_U$ can be useful for direct interpretations of DIF in terms of differing expectations of total score for the two groups.

In simulation studies presented here d_T was taken as zero. That is, the difference between the target ability means in the two groups was zero[3]. For simulation studies where

$d_T \neq 0$, see Shealy & Stout (1992b). Two values of C_β were considered: 0.5, and 1.0. Positive values of C_β denote DIF against the focal group and negative values of C_β denote DIF against the reference group. Three different combinations of examinee sizes (J_F, J_R), typical of those commonly occurring in applications, were considered: $(J_F=500,\ J_F=500)$, (1000, 3000), and (1000, 1000). Two valid subtest lengths (N) were considered: 25 and 50 items. These items were randomly selected from 80 estimated three–parameter logistic item parameters. Item responses for the valid subtest were generated by using the three–parameter logistic model:

$$P_i(\theta) = c_i + \frac{1-c_i}{1+exp(-1.7(a_i(\theta-b_i))}, \ i=1,\ldots,n \qquad (8)$$

where a_i, b_i, and c_i are the discrimination, difficulty and guessing parameters of item i. Item responses for the studied subtest were generated by using the two–dimensional three parameter logistic model with compensatory abilities (Reckase & McKinley, 1983):

$$P_i(\theta,\eta) = c_i + \frac{1-c_i}{1+exp(-1.7(a_{i\theta}(\theta-b_{i\theta})+a_{i\eta}(\eta-b_{i\eta})))}, \ i=n+1,\ldots,N \qquad (9)$$

For each simulated examinee (see Equation 2), binary item responses (0,1) were obtained as follows. The probability of correctly answering valid subtest items was computed using Equation 8. If a simulated uniform random value on the interval (0,1) was less than or equal to the computed $P_i(\theta)$, then the item was considered answered correctly and a score of 1 was assigned. Otherwise the item was considered incorrect and a score of 0 was assigned. Similarly, for studied items $P_i(\theta,\eta)$ was computed using Equation 9 and a score value of 0 or 1 was assigned.

Cancellation Study

Since there are two nuisance abilities η_1 and η_2 in this case, these are generated as follows. The θ and η_1 have a bivariate normal distribution given by

$$\begin{pmatrix} \Theta & | g \\ \eta_1 & | g \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\theta g} \\ \mu_{\eta_1 g} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], \qquad (10)$$

and θ and η_2 have a bivariate normal distribution given by

$$\begin{pmatrix} \Theta & | g \\ \eta_2 & | g \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\theta g} \\ \mu_{\eta_2 g} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right] \qquad (11)$$

where ρ is the correlation between θ and η_1, and between θ and η_2, which is taken to be 0.5 for both groups. Also, η_1 and η_2 were generated independently of each other, for each fixed θ. As in the case of amplification, variances $\sigma^2(\theta | g)$, $\sigma^2(\eta_1 | g)$ and $\sigma^2(\eta_2 | g)$ were all taken to be 1. Thr means $\mu_{\theta g}$ ($\mu_{\theta R}$ and $\mu_{\theta F}$) were determined by Equations 3 and 4. The means $\mu_{\eta_1 g}$ ($\mu_{\eta_1 R}$ and $\mu_{\eta_1 F}$) and $\mu_{\eta_2 g}$ ($\mu_{\eta_2 R}$ and $\mu_{\eta_2 F}$) were determined through Equations 12 and 13 as follows:

$$C_{\beta i} = E[\eta_{iR} | \theta] - E[\eta_{iF} | \theta]$$

$$= (\mu_{\eta_i R} - \mu_{\eta_i F}) - \rho d_T, \qquad i=1,2 \qquad (12)$$

and

$$(\mu_{\eta_i R} - \rho\mu_{\theta R}) = -(\mu_{\eta_i F} - \rho\mu_{\theta F}), \qquad i=1,2 \qquad (13)$$

where $C_{\beta i}$ is the potential for DIF caused by the nuisance ability η_i and is chosen just as for the amplification case. Item responses were generated just as in the amplification case using Equations 8 and 9. Here Equation 9 applies to (θ,η_1) or (θ,η_2) depending upon item number. For example, items 1 through 11 of Table 1 depend upon θ and η_1, and items 12 through 14 depend upon θ and η_2.

Results of Simulation Study

Three different simulation studies were done, each with varying values for (J_R,J_F), C_β, and N. The results for Amplification Study 1 are shown in Table 2. This study has 500 examinees in each of the focal and reference groups with 50 items in the valid subtest. The first column denotes the item numbers (taken from Table 1) used in the studied subtest; the second column denotes the degree of potential for DIF induced in the simulations (C_β); the third column denotes the average estimated DIF over 100 replications $(\bar{\hat{\beta}}_{\mathrm{U}})$; the fourth column denotes the observed (estimated) standard error of $\hat{\beta}_{\mathrm{U}}$ over 100 replications; and the fifth column denotes the rejection rate of testing the null hypothesis of no DIF over 100 replications. The last three columns report the estimated mean, standard error, and the rejection rate of DIF using the Mantel–Haenszel statistic over 100 replications. The first row of Table 2, for example, denotes that item 4, from Table 1, was used in the studied subtest with .50 as the potential for DIF. The average amount of estimated DIF, over 100 replications, was .022 with a standard error of .036. The null hypothesis of no DIF was rejected 18 out of 100 replications. The Mantel–Haenszel analyses indicate that for this item, the estimated mean of Δ_{MH} was –.342 with an observed standard error of .435. The null hypothesis of no DIF was rejected 9 times out of 100 replications.

As can be seen from Table 2, each of the items 4, 5, 6, 7, and 8 were tested individually for DIF, and then tested collectively. That is, in each case the valid subtest consisted of 50 items and the studied subtest consisted of exactly one item except for the last row where the studied subtest consisted of all five items. It can be seen that the average amount of estimated DIF for individual items ranged from .022 to .035, indicating small DIF ($0 \leq \hat{\beta}_U < .05$) at the item level. When all five DIF items were included in the studied subtest, however, the amount of estimated DIF was amplified to .148, indicating a large DIF ($\hat{\beta}_U \geq .1$). In other words, when all DIF items act in concert, the difference in the expected test scores between the groups was about .15. Thus, from column three, it can be seen that at the item level each of these items are likely to be missed as DIF items because of their low value of estimated DIF, nonetheless, at the test level the amplification is such that the total DIF is substantial. Similarly from column five it can be seen that the rejection rate for individual items ranged from .17 to .23 while the rejection rate for all five items together jumped to .7, reflecting the cumulative effect of DIF. Comparison of SIBTEST results with those of Mantel–Haenszel show that both the procedures are consistent in their assessment of direction of DIF, the amount of estimated DIF, and the standard error of estimate, whenever a single item was considered.

Table 3 displays the results of Amplification Study 2. In this case the degree of potential for DIF was increased to 1.0 and the sample sizes for reference and focal groups were increased to 3000 and 1000 respectively. Items 9, 10, and 11 (from Table 1) were selected for this study. Similar to the results in Table 2, for individual DIF items, the amount of estimated DIF was moderate ($.05 \leq \hat{\beta}_U < .1$). However, when all three DIF items were included in the studied subtest, the amount of estimated DIF was amplified to .225, indicating large DIF. That is, when all three DIF items act in concert, the estimated difference in the expected test score between the groups was beyond 0.2. Comparison of results of SIBTEST with those of Mantel–Haenszel again showed that they are consistent and comparable whenever a single item was considered for DIF.

Table 4 displays the results of the Amplification and Cancellation Study. Each of the reference and focal groups contains 1000 examinees with 25 items in the valid subtest. Items 1, 2, and 3, which depend upon θ and η_1 were used here with 0.5 as the potential for DIF against the focal group ($C_{\beta 1}$ positive). These studied items were tested individually and collectively for DIF against the focal group. Items 12, 13, and 14, which depend upon θ and η_2 were used with –0.5 as the potential for DIF, but against the reference group ($C_{\beta 2}$ negative). These items were also studied individually and collectively for DIF against the reference group. Finally, all six items were used collectively with their corresponding positive and negative DIFs to study DIF cancellation. As can be seen from Table 4, items 1, 2, and 3 together exhibit large positive DIF against the focal group ($\hat{\beta}_U$=.188); while items 12, 13, and 14 exhibit large negative DIF against the reference group ($\hat{\beta}_U$=–.185); However, when items 1, 2, 3, 12, 13, and 14, were combined together in the studied subtest, the DIF canceled out at the test score level ($\hat{\beta}_U$=–.002). Thus, this test, in spite of having six DIF items, displays virtually no DIF at the test level. Note that SIBTEST was used both to detect the amplification of positive DIF for items 1, 2, and 3 and the amplification of negative DIF for items 12, 13, and 14, as well as the cancellation resulting from the combined influence of all six studied items.

In summary, the simulation studies have demonstrated the effectiveness of SIBTEST in detecting DIF amplification and DIF cancellation. This was established for different sample sizes and test lengths. Comparison of SIBTEST results with those of Mantel–Haenszel, at the item level, show that both are performing about equally well.

Real Data Study

Description of the Data

Three real data sets were used to investigate the effectiveness of SIBTEST to detect amplification and cancellation of DIF in a real application. The data sets considered were: the American College Testing program (ACT) mathematics test data, Form 39B, for males and females; The National Assessment of Educational Progress (NAEP), 1986 history test data for males and females, and for Blacks and Whites (NAEP, 1988). The mathematics data consists of 60 items with 2115 males and 2885 females. The history data consists of 36 items with 1225 males, 1215 females, 1711 Whites, and 447 Blacks. The analyses were carried out in the following manner.

For each of the data sets, DIF/DTF analyses were performed. That is, each item was analyzed for DIF with the rest of the items forming the "valid subtest". In the first stage of item level analyses, both SIBTEST and Mantel–Haenszel statistics were computed and compared for each item. In the second stage of test level analyses, items that exhibited moderate to large DIF according to both procedures were analyzed together to investigate DIF amplification and cancellation. For these analyses, each studied subtest consisted of a collection of items of one of three types: items favoring the focal group, or items favoring the reference group, or item favoring both groups (that is, some items favoring the reference group and other items favoring the focal group). Thus an attempt was made to study both amplification and cancellation, from the DTF perspective.

Results of Real Data Study

The results of the analyses of mathematics data for males and females are shown in Tables 5 and 6. Table 5 shows the results of individual item analyses (that is DIF analyses). The items listed were identified as exhibiting DIF by both the procedures, the SIBTEST and the Mantel–Haenszel[4]. The first half of Table 5 shows items exhibiting moderate ($.05 \leq \hat{\beta}_U < .1$) to large ($\hat{\beta}_U \geq .1$) amount of DIF favoring males. That is, these items are showing DIF against females. The second half of Table 5 shows items exhibiting moderate to large amount of DIF against males.

Table 6 shows DIF amplification and cancellation effects for items shown in Table 5. Table 6 shows items used in the studied subtest; whether studied items favor males or females; the amount of estimated DIF ($\hat{\beta}_U$); the value of the Shealy–Stout statistic (B of Equation 1) and the associated p–value. The first row of Table 6 shows DIF cancellation effect of items 17 and 19 together. Item 17 favors males with large DIF while item 19 favors females with large DIF, each at the item level. When these items were combined together, however, the DIF canceled out completely at the test level ($\hat{\beta}_U$=–.0006). That is, although each of the items is favoring a different group at the item level, together at the test level the DIF canceled out resulting in no difference in the expected test scores of the two groups. The second row of Table 6 shows DIF amplification of items showing moderate DIF, each against females at the item level. The third row shows DIF amplification of items showing moderate DIF, each against males at the item level. The last row shows DIF amplification and cancellation when all items favoring males (with moderate and large DIF) and all items favoring females are analyzed together. Because DIF amplification for items favoring only males is higher in magnitude than DIF amplification for items favoring only females, when all DIF items were combined, positive and negative DIF is not totally canceled out. That is, there is some overall DTF for these items against females ($\hat{\beta}_U$= .294).

Tables 7 and 8 show the results of the analyses of the history test for males and females. Analogous to Table 5, Table 7 shows items exhibiting moderate to large amounts of DIF, by both procedures, for both groups. Table 8 shows the results of DIF amplification and cancellation effects. In Table 7 there is only one item with large DIF favoring males. The rest of items exhibit moderate DIF. Therefore, Table 8 shows DIF amplification results for items favoring males only; amplification results for items favoring females only; and amplification and cancellation results for all DIF items. As can be seen from the last row of Table 8, there is almost total cancellation of DIF ($\hat{\beta}_U$=.018) when all DIF items were assessed together. Thus, there is no DTF present in this case.

Tables 9 and 10 show the results of the analyses of the history test for Whites and Blacks. Analogous to the above two cases, Table 9 shows DIF results at the item level and Table 10 shows DTF results at the test level. It can be seen from Table 9 that very few items favor Blacks relative to the number of items that favor Whites. Therefore Table 10 only contains amplification results for items favoring Whites only and amplification and cancellation results for all the DIF items from Table 9. As expected, in this case, the magnitude of DIF amplification against Blacks is large, and when all DIF items were combined together there is only moderate DIF cancellation with overall DTF remaining against Blacks.

In summary, findings of real data studies have replicated findings from simulated studies in the sense that both amplification and cancellation were established. The results of SIBTEST analyses at the item level were almost totally consistent with those of the Mantel–Haenszel both in the direction and the amount of estimated DIF. The amplification and cancellation results using SIBTEST with real data have demonstrated the capability of SIBTEST to address these issues in real settings. It should be emphasized that the real data studies were DIF/DTF and not bias studies. These results are encouraging for future applications of SIBTEST for studying the cumulative effects of DIF at the test score level.

For all three sets of real data, content analyses of DIF items were performed in an attempt to identify the possible correlates to the occurrence of DIF and DTF. Upon studying the mathematics items shown in Tables 5 and 6, it was found that items that favored males and displayed amplification required analytical/geometry knowledge, such as, properties of triangles and trapezoids, angles in a circle, volume of a box, etc.; whereas items that favored females and displayed amplification required computational knowledge such as factorization, solving equations, etc. Based on these informal content analyses of the two sets of items displaying amplification, one could cautiously conjecture that math education of males may tend to develop understanding of analytical concepts while math education of females may tend to develop computational skills. Similar conclusions were drawn by Drasgow (1987) about the content of biased items of a different version of the ACT mathematics test.

Similarly, the analyses of the history items for the male, female comparison revealed that items favoring males involved factual knowledge, such as location of different countries on the world map, dates of certain historical events, etc., whereas, items favoring females involved reasoning ability about the constitution, entrance to the League of Nations, etc.

Content analyses of history items for Blacks and Whites again revealed factual knowledge items favoring Whites. That is, these items required knowledge of the location of different countries on the world map, facts about World War II, etc. There were only three items that favored Blacks and a common secondary trait in these three items was not evident. It was also interesting to note that, across the three data sets, the difficulty level of items that exhibited DIF did not differ significantly from the difficulty level of the rest of the items in the respective tests. In other words DIF was not related to difficulty level of items.

Summary and Discussion

This paper has investigated DIF amplification and cancellation at the test score level and SIBTEST's ability to detect and estimate each. Based on simulation as well as real data analyses, SIBTEST demonstrated its effectiveness to assess DIF at the item level as well as at the test score level. As demonstrated, at the test score level the cumulative effect of DIF could either amplify or cancel out partially or completely. In addition, at the item level of analysis, comparison of SIBTEST with Mantel–Haenszel showed mutual consistency.

If one wants to detect bias rather than merely detect DIF or DTF, one of the requirements of SIBTEST is that it requires a valid subtest, which serves as an internally valid benchmark to assess bias against. On the face of it, this requirement may sound unrealistic. However, attempts by Ackerman (1991a, 1991b) and others seem promising in obtaining an empirically validated valid subtest that could greatly assist in bias analyses. As an alternative to using the "valid" subtest to match examinees, one could also use an external criterion of the intended to be measured ability in concert with or instead of the valid subtest.

Study of DIF at the item level as well as at the test level can be very useful for test construction purposes. It is well known that item responses are multiply determined in the sense that multiple traits determine an examinee's response to each item. The decision to remove/add items should not be based at the item level analyses alone but should consider the effect of such items at the test level. It is possible one could add/remove items in order to balance the influence of one or more of secondary traits. Moreover, since decisions about individuals are made at the test score level, it is important to simultaneously assess the cumulative effect of several DIF items affecting different subpopulations at the test score level. As emphasized by other researchers (Drasgow, 1987; Humphreys, 1986; Roznowski, 19897; Reith & Roznowski, 1991), inclusion of items with multiple determinants could significantly improve the predictive as well as the construct validity of a test. Based on the analyses presented herein, SIBTEST could greatly aid in this process.

Although a statistical hypothesis testing procedure can be useful in the detection of test bias or DTF, it is important to distinguish between statistically significant DTF from a practically significant amount of DTF. This is because with any statistical procedure, it is well known that with large sample sizes small differences in group performance can result in a statistically significant result. For example, Drasgow (1987) has shown, through Lord's chi–square's method, that a large significant chi–square statistic may only reflect moderate bias at the test score level, even when one third of the items are biased. In the present study, for example, it would be useful to know the practical significance of observing a $\hat{\beta}_U$ value of .1, .5, 1.0 etc. at the test score level. The estimated index of DIF, $\hat{\beta}_U$, should be useful in assessing whether the amount of DIF present is of practical importance.

SIBTEST although derived using IRT, uses simple means and variances of scores on valid and studied subtests to obtain test statistics. It is computationally simple and does not involve IRT parameter estimation, thereby avoiding estimation problems. Simulation and real data studies of this paper have demonstrated SIBTEST's potential for assessing amplification and cancellation of DIF in a variety of situations. Nonetheless, more studies with varied sample sizes, test sizes, and in diverse contexts would be useful to further establish its empirical utility. Menu driven code and a user's manual are available on request for interested users.

Notes

[1] If $P(\underline{\theta})$ denotes the item characteristic curve then the marginal $P(\theta)$ is gotten by integrating out $\underline{\eta}$ from $P(\theta,\underline{\eta})$ using the conditional density $f(\underline{\eta}|\theta)$. $P(\theta)$ is interpreted as the probability of a randomly chosen examinee with target ability θ getting the item right.

[2] For some applications, it can make more sense to use reference group examinees or the entire group of examinees.

[3] Generally one finds nonzero differences in group means on the target ability (that is, $d_T \neq 0$).. However, there are many realistic situations where no differences in group means exist. In the present study d_T was taken as zero mainly to keep the design simple. The effectiveness of SIBTEST to detect DIF for varying d_T values has been demonstrated by Shealy and Stout (1992b) and by Roussos (1992). In these studies d_T was used as a factor in the experimental design.

[4] Across the three data sets (total 132 items), there were seven items where there was inconsistency between the SIBTEST and the Mantel–Haenszel analyses. Three items exhibited DIF through SIBTEST only and four items exhibited DIF through Mantel–Haenszel only. These items were not included in the studied subtest.

REFERENCES

Ackerman, T. A. (1991a). A didactic explanation of item bias, item impact, and item validity .rom a multidimensional perspective. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Ackerman, T. A. (1991b). Measurement direction in a multidimensional latent space and the role it plays in bias detection. Paper presented at the 1991 International Symposium on Modern Theories in Measurement: Problems and Issues. Montebello, Quebec, Canada.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A Berk (Ed.), Handbook of methods for detecting test bias (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. Educational and Ps"chological Measurement. 28, 61–75.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel—Haenszel method. Applied Measurement in Education, 2, 217–233.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (RR–83–9). Princeton, NJ: Educational Testing Service.

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.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology. 72, 19–29.

Hambleton, R. K. & Rogers, H. J. (1989). Detecting potentially biased items: Comparison of IRT area and Mantel–Haenszel methods. Applied Measurement in Education, 2(4), 313–334.

Holland, P. W., & Thayer, D. T. (1988). Different:al item performance and the Mantel–Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), Test Validity (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates. 8, 173–181.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst (pp. 23–32). Seattle: University of Washington.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology. 71, 327–333.

Hunter, J. F. (1975). A critical analysis of the use of item means and item–test correlations to determine the presence or absence of content bias in achievement test items. A paper presented at the National Institute of Education Conference on Test Bias. Annapolis, MD.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 155–174). Vancouver, BC: Educational Research Institute of British Columbia.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine and J. Rost (Eds.), Latent trait and latent class models. (pp. 263–274). New York: Plenum Press.

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

NAEP (1988). National Assessment of Educational Progress 1985–86 public–use data tapes. Version 2.0. Users Guide. Educational Testing Service.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture–fair selection. Journal of Educational Measurement, 13, 3–29.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495–502.

Reith, J. & Roznowski, M. (1991). Predictive relations of tests containing differentially functioning items: Do biased items result in biased tests? Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Reckase, M. D., & McKinley, R. L. (1983). Some latent trait theory on a multidimensional latent space. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 199–227). Baltimore, MD: Johns Hopkins University Press.

Roussos, L. (1992). Effects of small sample size and studied–item parameters on SIBTEST and Mantel–Haenszel type–I error rates. Unpublished manuscript. University of Illinois, Champaign, Illinois.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. Journal of Applied Psychology, 72, 480–483.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143–152.

Shealy, R. & Stout, W. F. (1992a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. & Stout, W. F. (1992b). A model–based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. Psychometrika.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A., Camilli, G., & Averill, M. (1981) Comparisons of procedures for detecting test–item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317–375.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361–370.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika, 55, 293–326.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. Journal of Educational Measurement, 28, 197–219.

Table 1

Item Parameters of Studied Subtests for Simulation Studies

Item	$a_{i\theta}$	$b_{i\theta}$	$a_{i\eta_1}$	$b_{i\eta_1}$	$a_{i\eta_2}$	$b_{i\eta_2}$	c_i
1	1.0	0	0.80	0	0	0	.2
2	1.5	0	0.75	0	0	0	.2
3	2.0	0	1.00	0	0	0	.2
4	0.8	−.3	0.20	0	0	0	.2
5	1.0	0	0.25	0	0	0	.2
6	1.5	.3	0.35	0	0	0	.2
7	2.0	0	0.40	0	0	0	.2
8	1.2	0	0.30	0	0	0	.2
9	0.8	−.3	0.30	0	0	0	.2
10	1.0	0	0.40	0	0	0	.2
11	1.5	·3	0.50	0	0	0	.2
12	1.0	0	0	0	0.80	0	.2
13	1.5	0	0	0	0.75	0	.2
14	2.0	0	0	0	1.00	0	.2

Table 2

Amplification Study 1
J_F=500, J_R=500, N = 50, d_T=0, α=.05

	SIBTEST				Mantel–Haenszel		
Item	C_β	$\check{\beta}_U$	$SE(\hat{\beta}_U)$	Rejection rate	$\check{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$	Rejection rate
4	.50	.022	.036	.18	–.342	.435	.09
5	.50	.031	.031	.17	–.416	.398	.15
6	.50	.035	.035	.23	–.489	.423	.22
7	.50	.030	.039	.18	–.444	.450	.12
8	.50	.028	.039	.22	–.424	.445	.19
4,5,6, 7,8	.50	.148	.067	.70	–	–	–

Table 3

Amplification Study 2
J_F=1000, J_R=3000, N = 50, d_T=0, α=.05

	SIB				Mantel–Haenszel		
Item	C_β	$\check{\beta}_U$	$SE(\hat{\beta}_U)$	Rejection rate	$\check{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$	Rejection rate
9	1.0	.062	.015	.99	–.996	.223	1.00
10	1.0	.087	.019	1.00	–1.140	.272	1.00
11	1.0	.096	.019	1.00	–1.248	.256	1.00
9,10,11	1.0	.225	.028	1.00	–	–	–

Table 4

Amplification and Cancellation Study
J_F=1000, J_R=1000, N = 25, d_T=0, α=.05

Item	$C_{\beta 1}$	$C_{\beta 2}$	$\bar{\beta}_U$	$SE(\hat{\beta}_U)$	*Rejection rate*
1	0.5	–	.071	.021	.98
2	0.5	–	.060	.023	.90
3	0.5	–	.065	.021	.96
1,2,3	0.5	–	.188	.040	1.00
12	–	–0.5	–.074	.021	1.00
13	–	–0.5	–.058	.022	.82
14	–	–0.5	–.062	.021	.98
12,13,14	–	–0.5	–.185	.036	1.00
1,2,3 12,13,14	0.5	–0.5	–.002	.061	.02

Table 5

Results of Mathematics Test: Males vs Females
Item Level DIF Analyses: SIBTEST & Mantel–Haenszel[1]

Items favoring males		*Items favoring females*	
$.05 \le \hat{\beta}_U < .1$	$\hat{\beta}_U \ge .1$	$.1 < -\hat{\beta}_U \le .05$	$-\hat{\beta}_U \le .1$
23, 32, 34, 38 48, 52, 58	17	4, 5, 9, 14, 29	19

[1] *These items were identified as exhibiting DIF by both the SIBTEST and the Mantel–Haenszel*

Table 6

Results of Mathematics Test: Males vs Females
DTF Amplification and Cancellation: SIBTEST

items of the studied subtest	*favors males*	*favors females*	$\hat{\beta}_U$	*B*	*p*
17 & 19	–	–	–.0006	–.06	.524
23, 32, 34, 38, 48 52, 58	*yes*	–	0.523	12.85	.000
4, 5, 9, 14, 29	–	*yes*	–.340	–10.15	.000
22, 32, 34, 38, 48 52, 58, 17, 4, 5, 9 14, 29, 19	*yes*	–	0.294	4.68	.000

Table 7

Results of History Test: Males vs Females
Item Level DIF Analyses: SIBTEST & Mantel–Haenszel

Items favoring males		*Items favoring females*	
$.05 \leq \hat{\beta}_U < .1$	$\hat{\beta}_U \geq .1$	$.1 < -\hat{\beta}_U \leq .05$	$-\hat{\beta}_U \leq .1$
12, 15, 25, 30	1	9, 11, 22, 24, 34	–

Table 8

Results of History Test: Males vs Females
DIF Amplification and Cancellation: SIB

items of the studied subtest	*favors males*	*favors females*	$\hat{\beta}_U$	*B*	*p*
12, 15, 25, 30, 1	*yes*	–	0.437	9.02	.000
9, 11, 22, 24, 34	–	*yes*	–.381	–7.87	.000
12, 15, 25, 30, 1, 9, 11, 22, 24, 34	–	–	0.018	0.24	.405

Table 9

Results of History Test: Whites vs Blacks
Item Level DIF Analyses: SIBTEST & Mantel–Haenszel

Items favoring Whites		*Items favoring Blacks*	
$.05 \le \hat{\beta}_U < .1$	$\hat{\beta}_U \ge .1$	$.1 < -\hat{\beta}_U \le .05$	$-\hat{\beta}_U \le .1$
7, 11, 12, 16, 35	13, 14, 15 17, 32, 36	3, 4, 5	–

Table 10

Results of History Test: Whites vs Blacks
Item Level DIF Analyses: SIB

items of the studied subtest	*favors Whites*	*favors Blacks*	$\hat{\beta}_U$	*B*	*p*
all items favoring Whites only	*yes*	–	1.310	9.96	.000
all items favoring Whites and Blacks	*yes*	–	1.150	7.43	.000

Distribution List

Dr. Terry Ackerman
Educational Psychology
260C Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Terry Allard
Code 1142CS
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217-5000

Dr. Nancy Allen
Educational Testing Service
Princeton, NJ 08541

Dr. Gregory Anrig
Educational Testing Service
Princeton, NJ 08541

Dr. Phipps Arabie
Graduate School of Management
Rutgers University
92 New Street
Newark, NJ 07102-1895

Dr. Isaac I. Bejar
Law School Admissions
Services
Box 40
Newtown, PA 18940-0040

Dr. William O. Berry
Director of Life and
Environmental Sciences
AFOSR/NL, N1, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Thomas G. Bever
Department of Psychology
University of Rochester
River Station
Rochester, NY 14627

Dr. Menucha Birenbaum
Educational Testing
Service
Princeton, NJ 08541

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Dr. Gwyneth Boodoo
Educational Testing Service
Princeton, NJ 08541

Dr. Richard L. Branch
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Dr. Robert Brennan
American College Testing
Programs
P. O Box 168
Iowa City, IA 52243

Dr. David V. Budescu
Department of Psychology
University of Haifa
Mount Carmel, Haifa 31999
ISRAEL

Dr. Gregory Candell
CTB/MacMillan/McGraw-Hill
2500 Garden Road
Monterey, CA 93940

Dr. Paul R. Chatelier
Perceptronics
1911 North Ft. Myer Dr.
Suite 1100
Arlington, VA 22209

Dr. Susan Chipman
Cognitive Science Program
Office of Naval Research
800 North Quincy St.
Arlington, VA 22217-5000

Dr. Raymond E. Christal
UES LAMP Science Advisor
AL/HRMIL
Brooks AFB, TX 78235

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director
Life Sciences, Code 1142
Office of Naval Research
Arlington, VA 22217-5000

Commanding Officer
Naval Research Laboratory
Code 4827
Washington, DC 20375-5000

Dr. John M. Cornwell
Department of Psychology
I/O Psychology Program
Tulane University
New Orleans, LA 70118

Dr. William Crano
Department of Psychology
Texas A&M University
College Station, TX 77843

Dr. Linda Curran
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. Charles E. Davis
Educational Testing Service
Mail Stop 22-T
Princeton, NJ 08541

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 1230F
University of Maryland
College Park, MD 20742

Dr. Sharon Derry
Florida State University
Department of Psychology
Tallahassee, FL 32306

Hei-Ki Dong
Bellcore
6 Corporate Pl.
RM: PYA-1K207
P.O. Box 1320
Piscataway, NJ 08855-1320

Dr. Neil Dorans
Educational Testing Service
Princeton, NJ 08541

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
(2 Copies)

Dr. Richard Duran
Graduate School of Education
University of California
Santa Barbara, CA 93106

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Engelhard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

ERIC Facility-Acquisitions
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Marshall J. Farr
Farr-Sight Co.
2520 North Vernon Street
Arlington, VA 22207

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-HR
The Pentagon
Washington, DC 20310-0300

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Chair, Department of
Computer Science
George Mason University
Fairfax, VA 22030

Dr. Robert D. Gibbons
University of Illinois at Chicago
NPI 909A, M/C 913
912 South Wood Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Susan R. Goldman
Peabody College, Box 45
Vanderbilt University
Nashville, TN 37203

Dr. Timothy Goldsmith
Department of Psychology
University of New Mexico
Albuquerque, NM 87131

Dr. Sherrie Gott
AFHRL/MOMJ
Brooks AFB, TX 78235-5601

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305-3096

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Patrick R. Harrison
Computer Science Department
U.S. Naval Academy
Annapolis, MD 21402-5002

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 13
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Prof. Lutz F. Hornke
Institut fur Psychologie
RWTH Aachen
Jaegerstrasse 17/19
D-5100 Aachen
WEST GERMANY

Ms. Julia S. Hough
Cambridge University Press
40 West 20th Street
New York, NY 10011

Dr. William Howell
Chief Scientist
AFHRL/CA
Brooks AFB, TX 78235-5601

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Martin J. Ippel
Center for the Study of
Education and Instruction
Leiden University
P. O. Box 9555
2300 RB Leiden
THE NETHERLANDS

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Professor Douglas H. Jones
Graduate School of Management
Rutgers, The State University
of New Jersey
Newark, NJ 07102

Dr. Brian Junker
Carnegie-Mellon University
Department of Statistics
Pittsburgh, PA 15213

Dr. Marcel Just
Carnegie-Mellon University
Department of Psychology
Schenley Park
Pittsburgh, PA 15213

Dr. J. L. Kaiwi
Code 442/JK
Naval Ocean Systems Center
San Diego, CA 92152-5000

Dr. Michael Kaplan
Office of Basic Research
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Jeremy Kilpatrick
Department of
Mathematics Education
105 Aderhold Hall
University of Georgia
Athens, GA 30602

Ms. Hae-Rim Kim
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State
University
Murfreesboro, TN 37132

Dr. Sung-Hoon Kim
KEDI
92-6 Umyeon-Dong
Seocho-Gu
Seoul
SOUTH KOREA

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. James Kraatz
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. Patrick Kyllonen
AFHRL/MOEL
Brooks AFB, TX 78235

Ms. Carolyn Laney
1515 Spencerville Rod
Spencerville, MD 20868

Richard Lanterman
Commandant (G-PWP)
US Coast Guard
2100 Second St., SW
Washington, DC 20593-0001

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
1310 South Sixth Street
University of IL at
Urbana-Champaign
Champaign, IL 61820-6990

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Hsin-hung Li
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Library
Naval Training Systems Center
12350 Research Parkway
Orlando, FL 32826-3224

Dr. Marcia C. Linn
Graduate School
of Education, EMST
Tolman Hall
University of California
Berkeley, CA 94720

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Logicon Inc. (Attn: Library)
Tactical and Training Systems
Division
P.O. Box 85158
San Diego, CA 92138-5158

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Evans Mandes
George Mason University
4400 University Drive
Fairfax, VA 22030

Dr. Paul Mayberry
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. James R. McBride
HumRRO
6430 Elmhurst Drive
San Diego, CA 92120

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541

Dr. Joseph McLachlan
Navy Personnel Research
and Development Center
Code 14
San Diego, CA 92152-6800

Alan Mead
c/o Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Timothy Miller
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. Ivo Molenar
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The NETHERLANDS

Dr. E. Muraki
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Ratna Nandakumar
Educational Studies
Willard Hall, Room 213E
University of Delaware
Newark, DE 19716

Academic Progs. & Research Branch
Naval Technical Training Command
Code N-62
NAS Memphis (75)
Millington, TN 30854

Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Norman, OK 73071

Head, Personnel Systems Department
NPRDC (Code 12)
San Diego, CA 92152-6800

Director
Training Systems Department
NPRDC (Code 14)
San Diego, CA 92152-6800

Library, NPRDC
Code 041
San Diego, CA 92152-6800

Librarian
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
Washington, DC 20375-5000

Office of Naval Research,
Code 1142CS
800 N. Quincy Street
Arlington, VA 22217-5000
(6 Copies)

Special Assistant for Research
Management
Chief of Naval Personnel (PERS-01JT)
Department of the Navy
Washington, DC 20350-2000

Dr. Judith Orasanu
Mail Stop 239-1
NASA Ames Research Center
Moffett Field, CA 94035

Dr. Peter J. Pashley
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dept. of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Dr. Peter Pirolli
School of Education
University of California
Berkeley, CA 94720

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Mr. Steve Reise
Department of Psychology
University of California
Riverside, CA 92521

Mr. Louis Roussos
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Donald Rubin
Statistics Department
Science Center, Room 608
1 Oxford Street
Harvard University
Cambridge, MA 02138

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37966-0900

Dr. Mary Schratz
4100 Parkside
Carlsbad, CA 92008

Mr. Robert Semmes
N218 Elliott Hall
Department of Psychology
University of Minnesota
Minneapolis, MN 55455-0344

Dr. Valerie L. Shalin
Department of Industrial
Engineering
State University of New York
342 Lawrence D. Bell Hall
Buffalo, NY 14260

Mr. Richard J. Shavelson
Graduate School of Education
University of California
Santa Barbara, CA 93106

Ms. Kathleen Sheehan
Educational Testing Service
Princeton, NJ 08541

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. Randall Shumaker
Naval Research Laboratory
Code 5500
4555 Overlook Avenue, S.W.
Washington, DC 20375-5000

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. David Thissen
Psychometric Laboratory
CB# 3270, Davie Hall
University of North Carolina
Chapel Hill, NC 27599-3270

Mr. Thomas J. Thomas
Federal Express Corporation
Human Resource Development
3035 Director Row, Suite 501
Memphis, TN 38131

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Elizabeth Wald
Office of Naval Technology
Code 227
800 North Quincy Street
Arlington, VA 22217-5000

Dr. Michael T. Waller
University of
Wisconsin-Milwaukee
Educational Psychology Dept.
Box 413
Milwaukee, WI 53201

Dr. Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Thomas A. Warm
FAA Academy
P.O. Box 25082
Oklahoma City, OK 73125

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Douglas Wetzel
Code 15
Navy Personnel R&D Center
San Diego, CA 92152-6800

German Military
Representative
Personalstammamt
Koelner Str. 262
D-5000 Koeln 90
WEST GERMANY

Dr. David Wiley
School of Education
and Social Policy
Northwestern University
Evanston, IL 60208

Dr. Bruce Williams
Department of Educational
Psychology
University of Illinois
Urbana, IL 61801

Dr. Mark Wilson
School of Education
University of California
Berkeley, CA 94720

Dr. Eugene Winograd
Department of Psychology
Emory University
Atlanta, GA 30322

Dr. Martin F. Wiskoff
PERSEREC
99 Pacific St., Suite 4556
Monterey, CA 93940

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
03-0T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Ms. Duanli Yan
Educational Testing Service
Princeton, NJ 08541

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G Street, N.W.
Washington, DC 20550