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

Ratna Nandakumar
Department of Educational Studies
University of Delaware

August 15, 1992


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.
**Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF**

Ratna Nandakumar

Department of Statistics
University of Illinois
725 South Wright Street
Champaign, IL 61820

Cognitive Sciences Program
Office of Naval Research
800 N. Quincy
Arlington, VA 22217-5000

Approved for public release; distribution unlimited

To be published in Journal of Educational Measurement

See reverse
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 level 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 $\theta$ is the ability intended to be measured by the test, the nuisance determinant(s) $\eta$ 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),

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 framework 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 $\theta$. 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 \( \theta \) underlying the test items is assumed to be multidimensional: \( \{ \theta = (\theta, \eta) \} \), where \( \theta \) is the target ability, intended to be measured by the test, and \( \eta \) 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 \( \theta \). Knowledge of sports \( \eta_1 \) and knowledge of domestics \( \eta_2 \) are nuisance abilities. Let \( U \) denote the test response vector and \( h(U) \) 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 \( \theta \), and some items (biased items) measure both target ability and one or more nuisance abilities \( \eta \). It is also assumed that the usual IRT assumptions of local independence, monotonicity, and group invariance hold with respect to \( \theta \) and that this collection of assumptions do not hold for any subset of components of \( \theta \).

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:

\[
H_0 : \beta_U = 0 \quad \text{vs.} \quad H_1 : \beta_U > 0,
\]

where \( \beta_U \) 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 $\theta$. That is, marginal item characteristic curves\(^1\) for the two groups do not cross as $\theta$ varies over the ability range. Let $X = \Sigma^n U_i$ 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 $Y = \Sigma^n U_{i+1}$ 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 $\theta$ 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

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

(1)

where $\hat{\beta}_U = \sum^K_{0} p_k (\bar{Y}_{Rk} - \bar{Y}_{Fk})$, and $p_k$ 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 $X=k$ ($k=0,1,...,n$) 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 $\theta$ distributions between reference and focal groups arising from short test lengths (for example, 25 items), and inherent differences in the $\theta$ distributions for the two groups (for details, see regression correction in Shealy &
Amplification and Cancellation of DIF

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

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

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 \( \alpha \) if the value of \( B \) exceeds the upper \( 100(1-\alpha) \)th percentile point of the standard normal distribution. \( \hat{\beta}_U \) is also the statistic used to estimate the amount of unidirectional bias \( \beta_U \). 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 power.

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
Amplification and Cancellation of DIF

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 $\theta$, $\eta_1$ and $\eta_2$. All studied subtest items, except the last three, are influenced by $\theta$ and $\eta_1$. The last three items are influenced by $\theta$ and $\eta_2$. The guessing level is fixed to 0.2 for all items. For amplification studies, only items with nuisance ability $\eta_1$ were used. For the amplification and cancellation study, both, items with nuisance ability $\eta_1$, and items with nuisance ability $\eta_2$ 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 $\eta_1$ is dropped.

\[
\begin{pmatrix}
\Theta \mid g \\
\eta \mid g
\end{pmatrix}
- N
\begin{pmatrix}
(\mu_\theta g) \\
(\mu_\eta g)
\end{pmatrix},
\]

(2)

where $\rho$ is the correlation between $\theta$ and $\eta$ for group $g$, which is set at 0.5 for both groups (different values of $\rho$ across groups tends to produce bidirectional DIF). As can be seen the variances $\sigma^2(\theta \mid g)$ and $\sigma^2(\eta \mid g)$ 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

\[
d_T = \frac{\mu_\theta R - \mu_\theta F}{\sigma_\theta P},
\]

(3)

where
Amplification and Cancellation of DIF

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

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 (\( \bar{\theta} \)) of the valid subtest items was assumed equal to the average target ability pooled across groups:

\[ \bar{\theta} = E[\Theta] = \alpha_R \mu_{\theta R} + \alpha_F \mu_{\theta F}. \quad (4) \]

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 \( \bar{\theta} \), 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.

Potential for DIF \( C_\beta \) is defined as the difference between the conditional expectation of \( \eta \) for the two groups, given by

\[ C_\beta = E[\eta_R | \theta] - E[\eta_F | \theta] \]

\[ = (\mu_{\eta R} - \mu_{\eta F}) + (\rho \frac{\sigma_{\eta R}}{\sigma_{\theta R}})(\theta - \mu_{\theta R}) - (\rho \frac{\sigma_{\eta F}}{\sigma_{\theta F}})(\theta - \mu_{\theta F}) \]

Following Equation 2 and \( \rho \)

\[ C_\beta = (\mu_{\eta R} - \mu_{\eta F}) - \rho d_T \quad (5) \]
Another criterion for choosing the means of $\eta$ is that, for an "average" value of target ability ($\Theta=0$) 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,

$$(\mu_{\eta R} - \rho \mu_{\Theta R}) = -(\mu_{\eta F} - \rho \mu_{\Theta F}) \quad (6)$$

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

The choice of values for $C_\beta$ in the simulations were guided by the desired amount of the estimated DIF, $\hat{\beta}_U$. In other words, values of $C_\beta$ were chosen so that the amount of estimated DIF would be "small" ($0 \leq \hat{\beta}_U < 0.05$), "moderate" ($0.05 \leq \hat{\beta}_U < 0.1$), or "large" ($\hat{\beta}_U \geq 0.1$). 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 $\Delta_{MH}$ (Holland & Thayer, 1988). An approximate empirical relationship between $\Delta_{MH}$ and $\hat{\beta}_U$ is given by

$$\hat{\beta}_U \approx -\Delta_{MH}/10 \quad (7)$$

Recall that $\beta_U$ is a measure of the average difference in expected test scores between reference and focal group members of similar ability. That is, $\beta_U$ 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.
 Amplification and Cancellation of DIF

d_{\text{\beta}} \neq 0$, see Shealy & Stout (1992b). Two values of $C_{\beta}$ were considered: 0.5, and 1.0. Positive values of $C_{\beta}$ denote DIF against the focal group and negative values of $C_{\beta}$ 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: $(J_{F}=500, J_{R}=500)$, (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:

$$P_{i}(\theta) = c_{i} + \frac{1 - c_{i}}{1 + \exp(-1.7(a_{i}(\theta - b_{i}))}, i=1,...,n$$

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):

$$P_{i}(\theta, \eta) = c_{i} + \frac{1 - c_{i}}{1 + \exp(-1.7(a_{i}(\theta - b_{i}) + a_{i}(\eta - b_{i}))), i=n+1,...,N}$$

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 \( \eta_1 \) and \( \eta_2 \) in this case, these are generated as follows. The \( \theta \) and \( \eta_1 \) have a bivariate normal distribution given by

\[
\begin{pmatrix} \Theta | g \\ \eta_1 | g \end{pmatrix} - N\left[ \begin{pmatrix} \mu_{\theta | g} \\ \mu_{\eta_1 | g} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right],
\]

and \( \theta \) and \( \eta_2 \) have a bivariate normal distribution given by

\[
\begin{pmatrix} \Theta | g \\ \eta_2 | g \end{pmatrix} - N\left[ \begin{pmatrix} \mu_{\theta | g} \\ \mu_{\eta_2 | g} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]
\]

where \( \rho \) is the correlation between \( \theta \) and \( \eta_1 \), and between \( \theta \) and \( \eta_2 \), which is taken to be 0.5 for both groups. Also, \( \eta_1 \) and \( \eta_2 \) were generated independently of each other, for each fixed \( \theta \). 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. The 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:

\[
C_{\beta i} = E[\eta_i R | \theta] - E[\eta_i F | \theta]
\]

\[
= (\mu_{\eta_i R} - \mu_{\eta_i F}) - \rho d_{R i}, \quad i=1,2 \tag{12}
\]

and

\[
(\mu_{\eta_i R} - \rho \mu_{\theta R}) = -(\mu_{\eta_i F} - \rho \mu_{\theta F}), \quad i=1,2 \tag{13}
\]
where $C_{\beta_i}$ is the potential for DIF caused by the nuisance ability $\eta_i$ 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 $(\theta, \eta_1)$ or $(\theta, \eta_2)$ depending upon item number. For example, items 1 through 11 of Table 1 depend upon $\theta$ and $\eta_1$, and items 12 through 14 depend upon $\theta$ and $\eta_2$.

Results of Simulation Study

Three different simulation studies were done, each with varying values for $(J_R, J_F)$, $C_{\beta}$ 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 ($C_{\beta}$); the third column denotes the average estimated DIF over 100 replications ($\tilde{\beta}_U$); the fourth column denotes the observed (estimated) standard error of $\tilde{\beta}_U$ 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 $\Delta_{MH}$ 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
Amplification and Cancellation of DIF

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 ($0 < \beta_U < .05$) 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 ($\beta_U \geq .1$). 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 ($.05 < \beta_U < .1$). 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 $\theta$ and $\eta_1$ 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 $\theta$ and $\eta_2$ 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 ($\bar{\beta}_U = .188$); while items 12, 13, and 14 exhibit large negative DIF against the reference group ($\bar{\beta}_U = - .185$); 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 ($\bar{\beta}_U = - .002$). 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. The first half of Table 5 shows items exhibiting moderate ($0.05 \leq \hat{\beta}_U < 0.1$) to large ($\hat{\beta}_U \geq 0.1$) 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 ($\beta_{ij}=.018$) 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
Amplification and Cancellation 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 $\beta_U$ 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 $P(\theta)$ denotes the item characteristic curve then the marginal $P(\theta)$ is gotten by integrating out $\eta$ from $P(\theta,\eta)$ using the conditional density $f(\eta|\theta)$. $P(\theta)$ is interpreted as the probability of a randomly chosen examinee with target ability $\theta$ 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


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*.


<table>
<thead>
<tr>
<th>Item</th>
<th>$a_{i\theta}$</th>
<th>$b_{i\theta}$</th>
<th>$a_{i\eta_1}$</th>
<th>$b_{i\eta_1}$</th>
<th>$a_{i\eta_2}$</th>
<th>$b_{i\eta_2}$</th>
<th>$c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>0</td>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>0</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>-.3</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>.3</td>
<td>0.35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>7</td>
<td>2.0</td>
<td>0</td>
<td>0.40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>8</td>
<td>1.2</td>
<td>0</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>-.3</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>0</td>
<td>0.40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>11</td>
<td>1.5</td>
<td>.3</td>
<td>0.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.80</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>13</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>14</td>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>0</td>
<td>.2</td>
</tr>
</tbody>
</table>
Table 2
Amplification Study 1
$J_F=500$, $J_R=500$, $N = 50$, $d_T=0$, $\alpha=.05$

<table>
<thead>
<tr>
<th>Item</th>
<th>$C_\beta$</th>
<th>$\bar{\beta}_U$</th>
<th>$SE(\hat{\beta}_U)$</th>
<th>Rejection rate</th>
<th>$\bar{\Delta}_{MH}$</th>
<th>$SE(\hat{\Delta}_{MH})$</th>
<th>Rejection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.50</td>
<td>.022</td>
<td>.036</td>
<td>.18</td>
<td>-.342</td>
<td>.435</td>
<td>.09</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
<td>.031</td>
<td>.031</td>
<td>.17</td>
<td>-.416</td>
<td>.398</td>
<td>.15</td>
</tr>
<tr>
<td>6</td>
<td>.50</td>
<td>.035</td>
<td>.035</td>
<td>.23</td>
<td>-.489</td>
<td>.423</td>
<td>.22</td>
</tr>
<tr>
<td>7</td>
<td>.50</td>
<td>.030</td>
<td>.039</td>
<td>.18</td>
<td>-.444</td>
<td>.450</td>
<td>.12</td>
</tr>
<tr>
<td>8</td>
<td>.50</td>
<td>.028</td>
<td>.039</td>
<td>.22</td>
<td>-.424</td>
<td>.445</td>
<td>.19</td>
</tr>
<tr>
<td>4,5,6</td>
<td>.50</td>
<td>.148</td>
<td>.067</td>
<td>.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7,8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3
Amplification Study 2
$J_F=1000$, $J_R=3000$, $N = 50$, $d_T=0$, $\alpha=.05$

<table>
<thead>
<tr>
<th>Item</th>
<th>$C_\beta$</th>
<th>$\bar{\beta}_U$</th>
<th>$SE(\hat{\beta}_U)$</th>
<th>Rejection rate</th>
<th>$\bar{\Delta}_{MH}$</th>
<th>$SE(\hat{\Delta}_{MH})$</th>
<th>Rejection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1.0</td>
<td>.062</td>
<td>.015</td>
<td>.99</td>
<td>-1.996</td>
<td>.223</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>.087</td>
<td>.019</td>
<td>1.00</td>
<td>-1.140</td>
<td>.272</td>
<td>1.00</td>
</tr>
<tr>
<td>11</td>
<td>1.0</td>
<td>.096</td>
<td>.019</td>
<td>1.00</td>
<td>-1.248</td>
<td>.256</td>
<td>1.00</td>
</tr>
<tr>
<td>9,10,11</td>
<td>1.0</td>
<td>.225</td>
<td>.028</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4
Amplification and Cancellation Study
$J_F=1000$, $J_R=1000$, $N = 25$, $d_T=0$, $\alpha=.05$

<table>
<thead>
<tr>
<th>Item</th>
<th>$C_{\beta_1}$</th>
<th>$C_{\beta_2}$</th>
<th>$\bar{\beta}_v$</th>
<th>$SE(\beta_v)$</th>
<th>Rejection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>-</td>
<td>.071</td>
<td>.021</td>
<td>.98</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>-</td>
<td>.060</td>
<td>.023</td>
<td>.90</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>-</td>
<td>.065</td>
<td>.021</td>
<td>.96</td>
</tr>
<tr>
<td>1,2,3</td>
<td>0.5</td>
<td>-</td>
<td>.188</td>
<td>.040</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-0.5</td>
<td>-.074</td>
<td>.021</td>
<td>1.00</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-0.5</td>
<td>-.058</td>
<td>.022</td>
<td>.82</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-0.5</td>
<td>-.062</td>
<td>.021</td>
<td>.98</td>
</tr>
<tr>
<td>12,13,14</td>
<td>-</td>
<td>-0.5</td>
<td>-.185</td>
<td>.036</td>
<td>1.00</td>
</tr>
<tr>
<td>1,2,3</td>
<td>0.5</td>
<td>-0.5</td>
<td>-.002</td>
<td>.061</td>
<td>.02</td>
</tr>
<tr>
<td>12,13,14</td>
<td>-</td>
<td>-0.5</td>
<td>-.002</td>
<td>.061</td>
<td>.02</td>
</tr>
</tbody>
</table>
Table 5

Results of Mathematics Test: Males vs Females
Item Level DIF Analyses: SIBTEST & Mantel-Haenszel

<table>
<thead>
<tr>
<th>Items favoring males</th>
<th>Items favoring females</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05 ≤ ( \hat{\beta}_u &lt; .1 ) ( \hat{\beta}_u \geq .1 )</td>
<td>.1 &lt; (-\hat{\beta}_u) ≤ .05 (-\hat{\beta}_u) ≤ .1</td>
</tr>
<tr>
<td>23, 32, 34, 38, 48, 52, 58</td>
<td>4, 5, 9, 14, 29, 19</td>
</tr>
</tbody>
</table>

1These 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

<table>
<thead>
<tr>
<th>items of the studied subtest</th>
<th>favors males</th>
<th>favors females</th>
<th>( \hat{\beta}_u )</th>
<th>B</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>17 &amp; 19</td>
<td>-</td>
<td>-</td>
<td>-.0006</td>
<td>-.06</td>
<td>.524</td>
</tr>
<tr>
<td>23, 32, 34, 38, 48, 52, 58</td>
<td>yes</td>
<td>-</td>
<td>0.523</td>
<td>12.85</td>
<td>.000</td>
</tr>
<tr>
<td>4, 5, 9, 14, 29</td>
<td>-</td>
<td>yes</td>
<td>-.340</td>
<td>-10.15</td>
<td>.000</td>
</tr>
<tr>
<td>22, 32, 34, 38, 48, 52, 58, 17, 4, 5, 9</td>
<td>yes</td>
<td>-</td>
<td>0.294</td>
<td>4.68</td>
<td>.000</td>
</tr>
<tr>
<td>14, 29, 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 7

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

<table>
<thead>
<tr>
<th>Items favoring males</th>
<th>Items favoring females</th>
</tr>
</thead>
<tbody>
<tr>
<td>$.05 \leq \hat{\beta}_u &lt; .1 \quad \hat{\beta}_u \geq .1</td>
<td>.1 &lt; -\hat{\beta}_u \leq .05 \quad -\hat{\beta}_u \leq .1</td>
</tr>
<tr>
<td>12, 15, 25, 30, 1</td>
<td>9, 11, 22, 24, 34</td>
</tr>
</tbody>
</table>

### Table 8

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

<table>
<thead>
<tr>
<th>items of the studied subtest</th>
<th>favors males</th>
<th>favors females</th>
<th>$\hat{\beta}_u$</th>
<th>$B$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12, 15, 25, 30, 1, 9, 11, 22, 24, 34</td>
<td>yes</td>
<td>–</td>
<td>0.437</td>
<td>9.02</td>
<td>.000</td>
</tr>
<tr>
<td>12, 15, 25, 30, 1, 9, 11, 22, 24, 34</td>
<td>–</td>
<td>yes</td>
<td>-.381</td>
<td>-7.87</td>
<td>.000</td>
</tr>
<tr>
<td>12, 15, 25, 30, 1, 9, 11, 22, 24, 34</td>
<td>–</td>
<td>–</td>
<td>0.018</td>
<td>0.24</td>
<td>.405</td>
</tr>
</tbody>
</table>
Table 9

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

<table>
<thead>
<tr>
<th>Items favoring Whites</th>
<th>Items favoring Blacks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>.05 ≤ $\hat{\beta}_u &lt; .1$ $\hat{\beta}_u ≥ .1$</td>
<td></td>
</tr>
<tr>
<td>7, 11, 12, 16, 13, 14, 15</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td>35 17, 32, 36</td>
<td></td>
</tr>
</tbody>
</table>

Table 10

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

<table>
<thead>
<tr>
<th>items of the studied subtest</th>
<th>favors</th>
<th>$\hat{\beta}_u$</th>
<th>B</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>all items favoring Whites</td>
<td>yes</td>
<td>-</td>
<td>1.310</td>
<td>9.96</td>
</tr>
<tr>
<td>only</td>
<td></td>
<td>1.150</td>
<td>7.43</td>
<td>.000</td>
</tr>
<tr>
<td>all items favoring Whites</td>
<td>yes</td>
<td>-</td>
<td>1.150</td>
<td>7.43</td>
</tr>
<tr>
<td>and Blacks</td>
<td></td>
<td>1.150</td>
<td>7.43</td>
<td>.000</td>
</tr>
</tbody>
</table>