DOCUMENT RESUME

ED 371 038 TM 021 742

AUTHOR Nandakumar, Ratna
TITLE Development of a Valid Subtest for Assessment of DIF/Bias.
PUB DATE Apr 94
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; Computer Simulation; Construct Validity; *Educational Assessment; Equal Education; *Item Bias; Research Methodology; *Test Construction; Test Items; Test Results
IDENTIFIERS *DIMTEST (Computer Program); *Multidimensionality (Tests)

ABSTRACT

By definition, differential item functioning (DIF) refers to unequal probabilities of a correct response to a test item by examinees from two groups when controlled for their ability differences. Simulation results are presented for an attempt to purify a test by separating out multidimensional items under the assumption that the intent of the test constructors was to construct a unidimensional test for a given population. The procedure used to arrive at the purified and essentially unidimensional subtest used the multidimensional theory of DIF/bias proposed by Shealy and Stout (1993) and the statistical procedure DIMTEST for assessing essential unidimensionality. When applicable, the proposed methodology leads to a statistically validated construct valid subtest that can be used in the matching criterion for DIF/bias analysis. This methodology can be applied to an internal or external matching criterion. It is only applicable when the majority of test items are tapping the intended ability for a given population, while a few items are tapping other major abilities in addition to the intended ability. This method is not applicable when DIF/bias is pervasive. Ten tables summarize analysis results. (Contains 28 references.) (SLD)

Development of a Valid Subtest for Assessment of DIF/Bias

Ratna Nandakumar
Department of Educational Studies
University of Delaware

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 7, 1994.

Development of a Valid Subtest for Assessment of DIF/Bias

Presently study of differential item functioning (DIF) or item bias is one of the major areas of research in educational measurement. In recent years numerous DIF studies have been conducted. By definition, DIF refers to unequal probabilities of correct response to an item by examinees from two groups when controlled for their ability differences. The term DIF therefore refers to different statistical properties of an item in the two groups of interest after controlling for the group differences in their ability. DIF in this sense is linked to the construct of the test items. The term bias on the other hand cannot be so simply explained. More often it is linked to the fairness of the test. According to Angoff (1993), bias conveys a social meaning that goes beyond the statistical differences between the group performance. Many definitions of bias, both historical and modern, and their implications are described in detail by Camilli (1993). Of these definitions, one proposed by Jensen (1980) states:

> *In mathematical statistics, "bias" refers to a systematic under or overestimation of a population parameter by a statistic based on samples drawn from the population. In psychometrics, bias refers to systematic errors in the predictive validity or construct validity of test scores of individuals that are associated with the individual's group membership. "Bias" is a general term and is not limited to "culture bias." It can involve any type of group membership—race, social class, nationality, sex, religion, age. The assessment of bias is purely objective, empirical, statistical and quantitative matter entirely independent of subjective value judgements and ethical issues concerning fairness or unfairness of tests and uses to which they are put. Psychometric bias is a set of statistical attributes conjointly of a given test and two or more specific subpopulations.* (p. 375)

Jensen's definition of bias is consistent with the meaning of DIF that is currently in use. In other words, the psychometric definition of bias according to Jensen refers to the objective, empirical, and statistical matter that is independent of subjective value judgements about the use of tests and factors influencing test scores that are outside of test construct. Stout (Shealy & Stout, 1993b), however, differentiates between psychometric definitions of DIF and bias. According to Stout, DIF refers to a situation where the matching subtest used for matching examinees in the two groups of interest (in order to control for their ability differences) could be the total test score or any other valid test score where no strong claim is made about the construct validity of this matching criterion. Bias on the other hand, according to Stout, refers to a situation

where a claim is made about the construct validity of the matching criterion in the sense of it not contaminated by items measuring traits other than the content of interest. That is, the matching subtest is measuring exactly what the test is supposed to be measuring. In this sense bias is a special case of DIF.

In doing DIF studies, most often, the groups are controlled for the ability differences by matching examinees in the two groups using the total test score. Although the total test score provides a practical and convenient method for matching examinees, the serious shortcoming of using the total test score is that it could include DIF items, thus contaminating the matching criterion. Linn (1993) describes the best matching criterion as follows

> *The ideal matching variable would be a perfectly valid and reliable, unbiased measure of the developed ability the test is intended to measure. Such a measure is obviously unavailable and test scores are generally the closest approximation that is available.*

He goes on to say

> *... unidimensionality should be added to the list of characteristics for the ideal matching criterion. Otherwise, DIF may be associated with valid differences along different dimensions.*

Some researchers in the past have attempted to obtain an unbiased matching criterion (Clauser, Mazor, & Hambleton, 1993; Williams, 1993). Clauser et. al used a two-step process proposed by Holland and Thayer (1988) to "purify" the matching criterion internal to the test. Williams on the other hand, used an unbiased matching criterion external to the test in her study. Douglas, Li, and Stout (1994) have used hierarchical cluster analyses combined with expert opinion to select items that are potentially biased for DIF/bias analyses.

Shealy and Stout (1993a, 1993b) argue that bias can only manifest itself through multidimensional abilities being present. In other words, if item responses are essentially unidimensional, there is no bias. Consistent with this viewpoint, the purpose of the proposed research is to develop a methodology to establish a bias free matching subtest through the use of DIMTEST, a statistical procedure for assessing essential unidimensionality. Throughout, it is assumed that the intention of the test is to measure

a unidimensional construct, but possibly some item responses are also influenced by other abilities which could lead to differential performance among groups. The purpose of this study is to identify, through unidimensionality analysis, items that are influenced by other major abilities and remove them from the test. The remaining items are again assessed for unidimensionality. In addition, the degree of DIF/bias contributed by multidimensional items is assessed.

Throughout the paper we will be using the terminology of DIF/bias. If one agrees with Stout's notion of bias and DIF, the proposed methodology would lead to bias analyses because the matching subtest will be free of multidimensional items. On the other hand, if one believes that the term bias refers to factors beyond statistical differences between the groups on the item(s) of interest, the proposed methodology would lead to DIF analyses with a statistically and psychometrically valid matching criterion that is essentially unidimensional.

In what follows, the description of DIMTEST procedure for assessing essential unidimensionality, and the SIBTEST procedure for assessing DIF/bias will be described. The proposed methodology to develop a valid subtest that is free of DIF/bias using DIMTEST will then be described, followed by the simulation details and the results of simulation study and real data study.

The DIMTEST Procedure

The hypothesis to test for essential unidimensionality can be stated as

$$H_0 : d_E = 1 \; vs. \; H_1 : d_E > 1$$

where d_E denotes essential dimensionality.

DIMTEST (Stout, Nandakumar, Junker, Chang, & Steidinger, 1993) assesses essential unidimensionality of a given set of item responses (found as a result of administering a set of items to a group of examinees) by splitting the items into three subtests: two short subtests AT1 and AT2, and a large subtest PT. Items of AT1 are first selected so that they are dimensionally homogeneous. Items of AT2 are matched in difficulty to the items of AT1 and the rest of the items form the subtest PT. There are several methods to select AT1 items. Simple factor analysis can be used to select AT1 items. Using factor analysis, a small set of items with highest loadings of the same sign on the second factor are selected (Nandakumar & Stout, 1993; Stout, 1987). Expert opinion

is another method to select AT1 items. Based on experience, one can select a small set of items (not more than one-quarter of total items) believed to be measuring the same trait (Nandakumar, 1993a). Alternatively, one can use hierarchical cluster analysis to select AT1 items (Roussos, Stout, & Marden, 1993). In the present study factor analysis was used to select AT1 items.

If item responses were driven by an essentially unidimensional model, items of all the subtests (AT1, AT2, and PT) would be of similar dimension. On the contrary, if item responses were driven by a multidimensional model, items of AT1 will be dimensionally homogeneous and differ from the rest of the items in the dimensional structure.

Item responses of the subtest PT are used to group examinees into K subgroups. Item responses of the subtest AT1 are used to compute two variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$ within each subgroup and their difference is appropriately standardized and summed across subgroups to arrive at the statistic T_1 given by

$$T_1 = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \left[\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \right]$$

The $\hat{\sigma}_k^2$ is the variance estimate of the AT1 subtest among examinees in the subgroup k, and $\hat{\sigma}_{U,k}^2$ is the estimate of the "unidimensional" variance computed by summing the item variances of the subtest AT1. S_k is the appropriate standard error computed for the subgroup k (for details see Nandakumar & Stout, 1993).

Similarly item responses of AT2 are used to compute variance estimates within subgroups and their difference is standardized and summed across subgroups to arrive at the statistic T_2. Stout's statistic T to assess for essential unidimensionality is then given by

$$T = \frac{T_1 - T_2}{\sqrt{2}} \tag{1}$$

which follows the standard normal distribution when the null hypothesis of essential unidimensionality holds.

When the null hypothesis of essential unidimensionality holds, the two variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$ should be approximately equal resulting in a small value for T_1. Although this is generally true, in certain situations T_1 could be inflated under H_0 due to difficulty of items in AT1 or due to shortness of the subtest PT (Nandakumar, & Stout, 1993; Stout, 1987). In such situations the statistic T_2 was designed to correct T_1

for inflation due to statistical biases. Consequently, under H_0, T will be small leading to the tenability of H_0. On the other hand when H_1 holds, the difference in the variance estimates will be large leading to the rejection of H_0.

The performance of DIMTEST has been studied extensively through Monte Carlo simulations in various test settings by varying parameters such as test length, sample size, ICC type, correlation between abilities, and the degree of multidimensionality (Nandakumar and Stout, 1993; Nandakumar, 1991; Stout, 1987). It has also been studied for its performance on various real tests (Nandakumar, 1993a). It has been found that DIMTEST has maintained desirable type-I error with high power even when the correlation between abilities is as large as 0.7.

The SIBTEST Procedure

It is assumed that two groups of examinees, the reference group (R), and the focal group (F) are administered a given set of items. These items are assumed to be measuring an intended ability (proficiency) denoted by θ, also known as the *target ability.* Unintended abilities that influence item responses are denoted by η *s*, and are known as *nuisance abilities.* It is also assumed that most items are influenced by one dominant ability, namely the target ability, and that a few items are influenced by more than one dominant ability, namely the target ability and the nuisance ability (or abilities) [1]. Items influenced by more than one dominant ability are considered to be potentially causing DIF/bias. For example, in a mathematics reasoning test, some items could be tapping the knowledge or familiarity with the baseball game, while some other items could be tapping the knowledge of the chess game in addition to knowledge in mathematics. In this instance, there are two nuisance abilities, the knowledge of baseball game (η_1), and the knowledge of chess game (η_2), and the knowledge of mathematics is the target ability (θ).

The SIBTEST (Stout & Roussos, 1992) has the unique capability of assessing cumulative DIF/bias of several items. That is, it can assess the amount of DIF/bias contributed by several items and its impact on the total test score, as opposed to single item DIF/bias which is due to one item at a time. The cumulative DIF/bias will be referred to as the differential test functioning (DTF) or the test bias.

The statistical hypothesis for testing the absence of DIF bias can be stated as

[1] It is possible that there could be one or more minor abilities influencing relatively few items to a small degree. These minor abilities, however, are not counted in assessing dimensionality.

$$H_0 : \beta_U = 0 \; vs. \; H_1 : \beta_U \neq 0$$

where β_U is the parameter denoting the amount of unidirectional DIF/bias when a single item is considered or DTF/test bias when more than one item is considered.

The procedure to test for DIF/bias using SIBTEST involves splitting of test items into two subtests: a "valid subtest" and a "studied subtest". The studied subtest contains potentially biased item(s) while the valid subtest, often but not always, contains the rest of the items. The valid subtest functions as the matching criterion.

If one is testing for DIF/DTF, without claiming the construct validity of the matching subtest, then the studied subtest would contain one item(s) for which DIF is being investigated with the rest of the items constituting the valid subtest (as is usually done with the Mantel-Haenszel test). On the other hand, if one is testing for DIF/bias claiming the valid subtest to be unidimensional, then the studied subtest would contain potentially biased item(s) while the valid subtest contains items judged to be free of DIF/bias, and is also essentially unidimensional.

Let N denote the total number of items. Let items 1 ... n denote valid subtest items, and items n+1 ... N denote studied subtest items. Let U_i denote the response to the item i taking values 0 or 1. For each examinee, let $X = \sum_{i=0}^{n} U_i$ denote the total score on the valid subtest and $Y = \sum_{i=n+1}^{N} U_i$ denote the total score on the studied subtest. Examinees within the reference and the focal groups are grouped into K subgroups (maximum K=n+1) according to their score on the valid subtest. Examinees with the same score on the valid subtest are compared across the reference and the focal groups on their performance on the studied subtest items. The weighted mean difference (between the reference and the focal groups) on the studied subtest score across K subgroups is then given by

$$\hat{\beta}_U = \sum_{k=0}^{K} \hat{p}_k (\overline{Y}_R k - \overline{Y}_F k) \tag{2}$$

where $\hat{p}_k$ denotes the proportion of focal group examinees in the subgroup k [2]. Here the $\overline{Y}_R k$ and $\overline{Y}_F k$ denote the adjusted means of the studied subtest (adjusted for mean differences between the reference and focal groups) for examinees in the reference and the focal groups respectively for subgroup k. $\hat{\beta}_U$ can also be used as a statistic that estimates the amount of unidirectional DIF/bias or DTF/test bias. For example a $\hat{\beta}_U$ value of .1 denotes that the average difference between the focal and the reference groups on the studied subtest item(s) is one tenth of a point when controlled for their ability differences. Significant positive values of $\hat{\beta}_U$ denote DIF/bias against the focal group

[2] one could also pool examinees from both groups of the kth subgroup.

examinees and significant negative values of $\hat{\beta}_U$ denote DIF/bias against the reference group examinees. The test statistic for testing the null hypothesis of no DIF/bias or DTF/test bias is given by

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \tag{3}$$

where $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$ given by

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=1}^{K} \hat{p}_k^2 \left(\frac{1}{J_R k} \hat{\sigma}^2 (Y|k, R) + \frac{1}{J_F k} \hat{\sigma}^2 (Y|k, F) \right) \right)^{1/2}$$

The null hypothesis is rejected at error rate α if B exceeds the upper $100(1 - \alpha)$th percentile point of the standard normal distribution.

The performance of SIBTEST for detecting the unidirectional DIF/bias has been studied extensively through Monte Carlo simulations by various researchers on various factors: the type–I error rate, power, sample sizes, varying degrees of DIF/bias induced, and for varying degrees of ability differences between the groups. (Ackerman, 1993; Narayanan & Swaminathan, 1993, Shealy & Stout, 1993b; Roussos & Stout, 1993). It has been found that the test statistic B has good adherence for type–I error with high power. The SIBTEST procedure can be applied to sample sizes as small as 100 and has exhibited higher power than other procedures for unequal ability distributions between the reference and the focal groups.

The performance of SIBTEST to assess cumulative DIF/bias has been studied by Nandakumar (1993b) for both simulated and real data sets. In her study both the amplification of DIF/bias due to one nuisance ability and cancellation of DIF/bias, partially or fully, due to two nuisance abilities were studied. Douglas, Li, Roussos, and Stout (1994) have further established the usefulness of SIBTEST to detect differential testlet functioning. In addition, SIBTEST has been expanded to detect crossing DIF/bias (in addition to the uniform DIF/bias), and to deal with polytomous items (Chang, Mazzeo, & Roussos, 1993; Li & Stout, 1993).

Proposed Method to Arrive at a DIF/Bias Free Valid Subtest

The basic premise of these analyses rests on the assumption that for DIF/bias to manifest, item responses should be influenced by more than one dominant ability (or

proficiency). If the item responses can be approximated by an essentially unidimensional model (hopefully this is the intended dimension), then there is no room for either DIF/bias or DTF/test bias. When data exhibits more than one dominant dimension, then an attempt can be made as described below to possibly identify these items and remove them from the test. If the remaining items are essentially unidimensional, then these items can be used as the matching subtest for DIF/bias analyses. We have proposed below several steps to arrive at a DIF/bias free matching subtest if it is possible to obtain one.

The first step is to test for essential unidimensionality of item responses. If the null hypothesis of essential unidimensionality—one dominant dimension is rejected, it means that multidimensionality is present in the data. That is, more than one dominant ability influences at least some items. The second step is to identify items contributing to multidimensionality. Recall that in this study it is assumed that most items are influenced by the intended ability and that a few items are also influenced by other dominant abilities. In order to identify items contributing to multidimensionality, again DIMTEST can be used. When DIMTEST rejects the null hypothesis, it means that items of the subtest AT1 differ in dimensionality structure from the rest of the items. In addition, these items would have high second factor loadings (relative to others) and therefore would be part of AT1. It is also true that items of the subtest AT1 could contain essentially unidimensional items in addition to multidimensional items (Nandakumar, 1993a). Therefore by examining the items of the subtest AT1 one could find items potentially causing multidimensionality. In order to separate out items that are truly multidimensional, DIMTEST has to be run repeatedly on the given sample of item responses, each time randomly splitting the examinee sample into two subsamples, one for selecting items for the subtest AT1 through factor analyses, and the other for computing Stout's unidimensionality statistic T. Items that consistently appear in AT1 across different repetitions of DIMTEST are considered as multidimensional items and are removed from the test items (see Nandakumar, 1993a for examples). In the present study DIMTEST was repeated 100 times and items that occured most frequently in AT1 were removed from the test. The third step is to assess the remaining items for essential unidimensionality. If the test statistic T is not statistically significant, it is concluded that the remaining items are essentially unidimensional, otherwise step two has to be repeated. The fourth step is to assess the direction and the degree of DIF/bias contributed by each of the items identified as multidimensional using essentially unidimensional items for the matching criterion. Following DIF/bias analyses, one could also do DTF/test bias analyses to study the cumulative impact (either amplification of DIF/bias due to one nuisance ability or cancellation of DIF/bias, partially or fully, due to more than one nuisance ability) of several multidimensional items at the total test score.

It is possible that at step two, one cannot find a consistent pattern of AT1 items across repeated use of DIMTEST, in spite of the null hypotheses of essential unidimensionality being rejected. Several conclusions can be drawn when this happens. First of all, the assumption that only a few items are influenced by multidimensionality is not valid. Either multidimensionality is pervasive in the test, or there are could be many major nuisance dimensions influencing different sets of items, or one major nuisance dimension could be influencing many items, etc. In such a situation serious investigation of item content has to be sought by experts.

Details about Simulation Study

The simulation program SIBSIM developed by Stout and Roussos (1992) was used to generate examinee responses for the reference and the focal groups in the present study. The simulations were designed to reflect the real world test and examinee characteristics as closely as possible.

Three test lengths were considered: 30, 40, and 50. For each test length, three levels of DIF/bias contaminations were used: 10%, 15%, and 20%. The first three columns of Table 1 display, for each test length, the percentages and the number of DIF/bias items. For example, in a 40-item test with 15% DIF/bias items, there are 6 contaminated items. For each test length, Table 1 also displays, in the last two columns, the item numbers for the valid subtest and for the studied subtest. As can be seen, for all tests, the DIF/bias items are included at the end of the test. These levels of DIF/bias contamination were chosen because it has been found in applications that it is not uncommon to find up to 30% of DIF items (Clauser, Mazor, & Hambleton, 1993; Oshima & Miller, 1992; Hambleton & Rogers, 1989).

The valid subtest consists of unidimensional items, and the studied subtest consists of two-dimesional items. In the present study only one nuisance ability is considered, therefore all studied subtest items are influenced by the target ability (θ) and the same nuisance ability (η) (see Nandakumar, 1993b for examples where there could be two nuisance abilities). Estimates of item parameters for the SAT-Verbal test (Drasgow, 1987) were used for the valid subtest. The SAT-Verbal test consists of 80 items. The desired number of items, for each test length, were randomly selected from the 80 items for the valid subtest. The item parameters of the studied subtest were obtained as follows.

From Table 1 it can be seen that the maximum number of studied subtest items needed were 10. Therefore, 10 items were randomly selected from the 80 SAT-Verbal

items and the discrimination and the difficulty parameters of these 10 items were treated as the parameters associated with the target ability, θ. The discrimination and the difficulty parameters associated with the nuisance ability, η were manipulated so that they are about 80% of the strength of the parameters associated with the target ability. The difficulty and the discrimination parameters for the 10 studied subtest items are displayed in Table2 for both the θ and the η. The desired number of items for the studied subtest in each case were randomly selected from Table2, except in the case of 50 items with 20% DIF/bias, where all 10 items were used.

Two ability distributions were used. In the first case, the distance between the means of the ability distributions of the groups (focal and reference), which is denoted by d_T, is set to 0. In the second case the d_T is set to 0.5. That is, in the second case the means for the two groups differed by 0.5 of a standard deviation. Both the target and the nuisance ability distributions were normally distributed. The standard deviations of abilty distributions for both groups were set to 1. The nuisance ability was generated independent of the target ability and the standard deviation of the nuisance ability distribution was also set to 1.

The amount of induced DIF/bias was specified through the difference between the means of the conditional distribution of the nuisance ability for the two groups and is denoted by k_β, which was set to 0.5. In other words, the conditional distributions of the nuisance ability for the focal and the reference groups differ by 0.5 standard deviations at all levels of the target abilitiy. Positive values of k_β denote DIF/bias against the focal group and negative values of k_β denote DIF/bias against the reference group. Two levels of guessing were considered, c=0 and c=0.2.

In summary, four factors were used in this study: three levels of test length (30, 40, and 50); three levels of DIF/bias contamination (10%, 15%, and 20%); two ability distributions (d_T=0 and 0.5); and two levels of guessing (c=0, and 0.2). There were a total of 36 conditions. For all 36 conditions, the amount of induced DIF/bias (k_β) was set 0.5, and the sample size was fixed to 1000 examinees.

Item responses for the focal and the reference group examinees were simulated as follows. The responses for the valid subtest items were generated using the unidimensional three–parameter logistic model given by

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + exp(-1.7a_i(\theta - b_i))}, i = 1, \ldots, n \qquad (4)$$

where a, b, and c denote respectively the item discrimination, item difficulty, and item guessing level. The responses for the studied subtest items were generated using the two–

dimensional three–parameter logistic model with compensatory abilities due to Reckase and McKinley (1983) given by

$$P_i(\theta,\eta) = c_i + \frac{1-c_i}{1+exp(-1.7(a_{i,\theta}(\theta - b_{i,\theta}) + a_{i,\eta}(\eta - b_{i,\eta})))}, i = n+1,\ldots,N \quad (5)$$

For each simulated examinee, the probability of correct response for each item was computed using either Equation 4 (for a valid subtest item), or Equation 5 (for a studied subtest item). If the computed probability was greater than the uniform random variable generated from the interval (0,1), the item was considered answered correctly and a score of 1 was assigned. Otherwise a score of 0 was assigned.

After generating item responses for the reference and the focal group examinees with the specified amount of induced DIF/bias, item responses on all items for the two groups were combined to form a single data set. At this stage we pretend that we know nothing about the data set. First we do dimensionality analyses using DIMTEST. If the null hypothesis of essential unidimensionality is rejected, we follow the steps as outlined earlier to separate out items causing multidimensionality. The remaining items were again tested for unidimensionality. If the associated p–value is greater than .05 it is concluded that the item response data are essentially unidimensional. Using this as the valid subtest that is free of DIF/bias contamination, multidimensional items are tested each separately and collectively for DIF/bias or DTF/test bias using the program SIBTEST.

Results

The details of dimensionality analyses are shown in Tables 3-6. Table 3 contains results for all test lengths and for different percentages of DIF/bias contaminations for the case of c=0 and $d_T = 0$; Table 4 contains results for the case of c=0.2 and $d_T = 0$; Table 5 contains results for the case of c=0 and $d_T = 0.5$; and Table 6 contains results for the case of c=0.2 and $d_T = 0.5$. The first two columns of Table 3 contains the test length and the percent of DIF/biased items for each test legth. For each condition, DIMTEST rejected the null hupothesis of essetial unidimensionality when applied on the item responses of the combined sample of the reference and the focal groups (with an exception for the case of 30 items with 10% DIF/bias contamination), which implied that multidimensionality was present in the data. In order to identify items causing multidimensionality, DIMTEST was repeated 100 times and items that consistently appeared in the subtest AT1 were noted. The third column gives the mean (over 100

trials) of Stout's statistic T to test for essential unidimensionality. The fourth column gives the rejection rate over 100 trials, and the fifth column gives items that most frequently occurred in the subtest AT1. The next two columns give the mimimum and the maximum percent of occurance of any of the AT1 items listed in column five, over 100 trials, and over the first 25 trials. The last column gives items that were removed from the test based on DIMTEST analyses. For example, for a 40-item test with 15% of DIF/biased items, the average T was 6.8, leading to the 99% rejection of the null hypothesis of essential unidimensionality. Item numbers 35 to 40 appeared most frequently in the subtest AT1, and these items appeared a minimum of 97% and a maximum of 98% for 100 trials. Similar percentages are reported for the first 25 trials, which are 96 to 100. Because these items appeared so frequently in AT1, they were suspected of causing multidimensionality and were removed from the test (In real world situations one would also look at the content of such items). We will first discuss results for all conditions except the case of 30 items with 10% DIF/bias contamination. At the end we will discuss results for this case separately.

From Table3 it can be seen that for all conditions the rejection rates were high (97 to 100), except for the case of 50 items with 10% DIF/bias contamination, where the rejection rate was 61%. Also, for every condition, multidimesional items (items in the last column of Table1) appeared most frequently in the subtest AT1. In the case of 50 items with 10% DIF/bias contamination, however, one unidimensional item also appeared most frequently in AT1 (we are trying to understand the reason for this). These items were therefore removed from the test. Comparing the percent occurance of AT1 items for all 100 trails, and for first 25 trials, it can be seen that repeating DIMTEST 25 times is as good as repeating it 100 times. The results are almost the same in both cases. In the case of 50 items with 10% contamination, although the rejection rate was only 61%, multidimensional items occured most freqeuntly in AT1 and thus lead to removal of these items.

Observation of results in Tables 4-6 lead to similar or better results. For all conditions the DIMTEST rejection rate was very high with multidimensional items falling into AT1 most frequently, leading to the rejection of H_0. Subsequently, these items were removed from the test.

In the case of 30 items with 10% DIF/bias, there were only 3 contaminated items. The DIMTEST requires a minimum of 4 items in the subtest AT1. Therefore, inevitably, one other unidimensional item got selected into AT1. Since AT1 was short, and mixed with unidimensional and multidimensional items, the rejection rate dropped to very low. However, the interesting fact was that even here multidimensional items occurred most frequently in AT1 and were removed from the test. In some conditions with 30 items,

one unidimensional item also got deleted (item #5).

In summary, the results displayed in Tables 3–6 reveal that dimensionality analyses of multidimensional data most often lead to high rejection rates over repeated trials of DIMTEST. Moreover, multidimensional items formed part of the subtest AT1 so that these items could be separated from the test. This was also true in cases of low (34%) or moderate (61%) rejection rates. That is, even in these cases multidimensional items were the most frequently occurring items of AT1 and lead to removal of these items from the test (although one additional unidimensional item was also removed in these two cases).

After removing multidimensional items from the test, the remaining items, in all 36 conditions, were again assessed for essential unidimensionality. If the remainining items are essentially unidimensional, then this subtest is used as the matching criterion for assessing DTF/test bias contributed by multidimensional items in SIBTEST analyses. The results of these analyses are displayed in Tables 7-10. The conditions of these tables are analgous to Tables 3-6. For example, Table 7, analgous to Table 3, displays results for the case of c=0 and $d_T = 0$. The first half of Table 7 displays results for dimensionality analyses, and the second half of Table 7 displays results for DTF/test bias analyses.

Under dimensionality analyses, in Table 7, the first two columns denote the value of Stout's T–statistic and the associated p–value for initial dimensionality analyses. The third column denotes the items deleted from the test following initial dimensionality analyses. The fourth and the fifth columns denote the value of Stout's T–statistic and the associated p–value for dimensionality analyses after removing items causing multidimensionality. As can be seen, in all cases, after removing multidimensional items the remaining item responses do conform to unidimensional modeling as evidenced by the T– and p–values. That is, the remaining items can be treated as essentially unidimensional. The same is true for conditions presented in Tables 8 through 10.

Using the essentially unidimensional items as the matching criterion, DTF/test bias analyses were performed using SIBTEST. The results for all conditions are displayed in the second half of Tables 7–10. The multidimensional items (that were removed) formed the studied subtest, while the remaining items, confirmed by DIMTEST as unidimensional, formed the valid subtest. The third column, under DTF/test bias analyes in Table 7, denotes the cumulative amount of estimated DTF/test bias caused by multidimensional items together. For example, a $\hat{\beta}_U$ value of 0.55 denotes that multidimensional items together contribute to a difference of about a five tenths of a point in the expected test scores between the focal and the reference groups. In other words, the reference group on the average, scores .55 of a point higher than the focal group after adjusting

for their ability differences. The last two columns give the value of the statistic B, and the p–value associated with the null hypothesis of no DTF/test bias. It can be seen in Tables 7-10 that for all conditions, multidimensional items contributed to significant DTF/test bias as evidenced by the B– and the p–values. We also performed the single item DIF/bias analyses, in addition to DTF/test bias analuses, (not reported here) for a random sample of cases and found that they were statistically significant.

In summary, after removing items causing multidimensionality, the remaining items were found to be unidimensional. The unidimensional items were then used as the matching criterion for DTF/Test bias analyses. For all conditions, multidimensional items contributed to a significant amount of DTF/test bias.

Summary and Discussion

The results presented here are preliminary findings, in an attempt to purify a test by separating out multidimensional items under the assumption that the intention of the test was to construct a unidimensional test for a given population. The procedure suggested in this article to arrive at a purified, essentially unidimensional subtest uses the multidimensional theory of DIF/bias proposed by Shealy and Stout (1993a), and the statistical procedure DIMTEST for assessing essential unidimensionality. When applicable, the proposed methodology leads to a statistically validated construct valid subtest that can be used as the matching criterion for DIF/bias analyses. This methodology can be applied on either an internal matching criterion or on an external matching criterion.

Within the context of the simulation study presented, the results look optimistic. For most DIF/bias conditions studied, we were able to identify multidimensional items that were contributing to DIF/bias, through unidimensional analyses. After removing the multidimensional items, the remaining items when confirmed as essentially unidimensional, were used as the matching subtest to assess the degree of DIF/DTF or item bias/test bias of the multidimensional items.

The simulation study presented here is limited in scope. It could be further improved on several factors. For example, the data generated in this study was either strictly unidimensional (for the valid subtest item responses), or strictly two-dimensional (for the studied subtest item responses). It would be interesting to introduce minor dimensions, in addition to major dimensions, that affect a few items to a small degree (that is, making the data essentially unidimensional or essentially multidimensional) and study their impact on DIF/bias as well as on DTF/test bias. Another improvement is the

strength of item parameters of the nuisance ability in relation to the strength of item parameters of the target ability, which can be made a factor in the study. Also the correlation between the target and the nuisance abilities could be a factor. In addition, it will be interesting to have more than one nuisance ability to examine the cancellation effect of DIF/bias either partially or fully.

As emphasized throughout this article, the proposed methodology is only applicable in a situation where the majority of test items are tapping the intended ability for a given population, and that a few items are tapping other major abilities in addition to the intended ability. If the DIF/bias is pervasive in test items, the proposed methodology is not applicable. In this regard Linn's (1993) quote is very appropriate:

> *We should recognize that DIF techniques, whether based on Mantel-Haenszel, the standardization, or an item response theory (IRT) procedure, all suffer from the serious limitation that they can detect only items that show large differences in one direction or the other relative to the total set of items. One cannot expect DIF procedures to detect "pervasive bias."*

He goes on to say

> *The fact that DIF procedures cannot be expected to detect pervasive bias should not be interpreted to mean that the procedures are of no value. Rather, it suggests that alone they are insufficient.*

References

Ackerman. T. (1993, April). *A didactic example of the influence of conditioning on the complete latent ability space when performing DIF analyses.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. W. Holland and H.Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camilli, G. (1993). The case against item bias detection technique based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland and H.Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chang, H., Mazzeo, J., & Roussos, L. (1993, April). *Extension of Shealy-Stout's DIF procedure to polytomous scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Clauser, B., Mazor, K., & Hambleton, R. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279.

Douglas, J., Li, H., Roussos, L., & Stout, W. (1994). *Testlet bias: Identifying suspect testlets and assessing testlet bias.* Paper submitted for publication.

Drasgow, F. (1987). A study of measurement bias of two standard psychological tests. *Journal of Applied Psychology, 72*, 19-30.

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

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

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Li, H., & Stout, W. (1993, June). *A new procedure for detection of crossing DIF/bias.* Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Linn, R. (1993). The use of differential item functioning statistic: A discussion of current practice and future implications. In P. W. Holland and H.Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nandakumar, R. (1993a). Assessing dimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

Nandakumar, R. (1993b). Simualtaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R., & Stout, W. (1993). Refinements of Stouts's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41-68.

Narayanan, P., & Swaminathan, H. (1993, April). *Performance of the Mantel-Haenszel and simultaneous item bias procedure for detecting differential item funtioning.* Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Apllied Psychological Measurement, 16*, 237-248.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Roussos, L., & Stout, W.(1993, June). *Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance.* Paper presented at the annual meeting of the AERA, Atlanta, Georgia.

Roussos, L., Stout, W., & Marden, J. (1994). *Analysis of the multidimensional structure of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper submitted for publication.

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

Shealy, R., & Stout, W. (1993a). An item response theory model for test bias. In P. W. Holland and H.Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, 197-239.

Stout, W., & Roussos, L. (1992). SIBTEST User Manual.

Stout, W., & Roussos, L. (1992). SIMTEST Program.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimesionality. *Psychometrika, 52*, 589-617.

Stout, W., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1993). DIMTEST: A Fortran program for assessing latent trait essential unidimensionality. *Applied Psychological Measurement, 16*, 236.

Williams, V. (1993, April). *Bridging gap between DIF and item bias.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Table 1

Percentages and Item Numbers of Valid and Studied Subtests Used in Simulations

Test Length	Percent of DIF/Bias Items	Number of DIF/Bias Items	Valid Subtest Item Numbers	Studied Subtest Item Numbers
30	10	3	1 to 27	28 to 30
	15	5	1 to 25	26 to 30
	20	6	1 to 24	25 to 30
40	10	4	1 to 36	37 to 40
	15	6	1 to 34	35 to 40
	20	8	1 to 32	33 to 40
50	10	5	1 to 45	46 to 50
	15	8	1 to 42	43 to 50
	20	10	1 to 40	41 to 50

Table 2

Parameters of Studied Subtests Used in Simulations

Item Number	a_θ	b_θ	a_η	b_η
1	1.10	−0.70	0.88	−0.56
2	0.70	−0.60	0.56	−0.48
3	0.90	−0.40	0.72	−0.32
4	1.40	0.10	1.10	0.08
5	0.90	0.90	0.72	0.72
6	1.20	0.70	0.96	0.56
7	0.90	0.30	0.72	0.24
8	1.00	0.80	0.80	0.64
9	1.60	1.10	1.28	0.88
10	2.00	1.10	1.60	0.88

Table 3

Details of Dimensionality Analyses for 100 Trials

$c = 0$, $d_T = 0$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Percent of DIF/Bias Items	T	Rejection Rate	Most Frequent AT1 Items	Percent Occurance of AT1 Items		Items Removed
					100 Trials	25 Trials	
30	10	1.0	34	5, 28 to 30	38-50	40-52	28 to 30
	15	6.3	100	26 to 30	96-100	96-100	26 to 30
	20	7.4	100	25 to 30	96-100	100	25 to 30
40	10	3.9	97	37 to 40	96-100	100	37 to 40
	15	6.8	99	35 to 40	97-98	96-100	35 to 40
	20	7.9	100	33 to 40	97-100	96-100	33 to 40
50	10	2.8	61	25, 46 to 50	48-64	48-64	25, 46 to 50
	15	8.7	100	43 to 50	76-100	88-100	43 to 50
	20	9.5	100	41 to 50	61-91	76-100	41 to 50

c = Guessing level

d_T = The mean difference between focal and reference groups on the target ability

J_F = The sample size for the focal group

J_R = The sample size for the reference group

k_β = The degree of potential for bias (the mean difference between the two groups on the nuisance ability)

$\hat{\beta}_U$ = The estimated DTF/Test Bias

Table 4

Details of Dimensionality Analyses for 100 Trials

$c = 0.2$, $d_T = 0$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Percent of DIF/Bias Items	T	Rejection Rate	Most Frequent AT1 Items	Percent Occurance of AT1 Items		Items Removed
					100 Trials	25 Trials	
30	10	1.1	34	5, 28 to 30	33-52	20-68	28 to 30
	15	4.6	98	26 to 30	97-100	96-100	26 to 30
	20	5.5	100	25 to 30	95-100	100	25 to 30
40	10	3.8	97	37 to 40	94-96	100	37 to 40
	15	7.0	100	35 to 40	98-100	100	35 to 40
	20	7.7	100	33 to 40	96-100	100	33 to 40
50	10	8.3	100	46 to 50	100	100	46 to 50
	15	10.0	100	43 to 50	99-100	100	43 to 50
	20	9.7	100	41 to 50	96-100	96-100	41 to 50

Table 5

Details of Dimensionality Analyses for 100 Trials

$c = 0$, $d_T = 0.5$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Percent of DIF/Bias Items	Rejection Rate	Most Frequent AT1 Items	Percent Occurance of AT1 Items		Items Removed
				100 Trials	25 Trials	
30	10	33	28 to 30	42	48	28 to 30
	15	100	26 to 30	91-100	92-100	26 to 30
	20	100	25 to 30	85-99	80-100	25 to 30
40	10	93	37 to 40	88-98	92-100	37 to 40
	15	98	35 to 40	87-97	84-100	35 to 40
	20	100	33 to 40	71-100	72-100	33 to 40
50	10	100	46 to 50	100	100	46 to 50
	15	100	43 to 50	99-100	96-100	43 to 50
	20	100	41 to 50	96-100	80-100	41 to 50

Table 6

Details of Dimensionality Analyses for 100 Trials

$c = 0.2$, $d_T = 0.5$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Percent of DIF/Bias Items	Rejection Rate	Most Frequent AT1 Items	Percent Occurance of AT1 Items		Items Removed
				100 Trials	25 Trials	
30	10	35	28 to 30	47-48	40-44	28 to 30
	15	99	26 to 30	92-99	92-100	26 to 30
	20	100	25 to 30	93-100	96-100	25 to 30
40	10	98	37 to 40	95-99	96	37 to 40
	15	100	35 to 40	75-100	72-100	35 to 40
	20	100	33 to 40	77-100	68-100	33 to 40
50	10	100	46 to 50	99-100	100	46 to 50
	15	100	43 to 50	96-100	96-100	43 to 50
	20	100	41 to 50	96-100	92-100	41 to 50

Table 7

Summary of Dimensionality and DTF/Test Bias Analyses

$c = 0$, $d_T = 0$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Dimensionality Analyses					DTF/Test Bias Analyses				
	T	p	Items Removed	T	p	Valid Subtest Length	Studied Subtest Length	$\hat{\beta}_U$	B	p
30	1.0	< .16	28-30	1.2	< .11	27	3	0.24	6.8	< 00
	6.3	< 00	26-30	0.4	< .32	25	5	0.55	10.3	< 00
	7.4	< 00	25-30	0.9	< .17	24	6	0.62	9.8	< 00
40	3.9	< 00	37-40	0.4	< .33	36	4	0.25	5.7	< 00
	6.8	< 00	35-40	−1.5	< .94	34	6	0.41	7.2	< 00
	7.9	< 00	33-40	−0.3	< .63	32	8	0.59	7.9	< 00
50	2.8	< 00	25, 46-50	−1.4	< .91	45	6	0.40	7.9	< 00
	8.7	< 00	43-50	0.3	< .38	42	8	0.63	8.7	< 00
	9.5	< 00	41-50	−1.6	< .95	40	10	0.78	8.5	< 00

c = Guessing level

d_T = The mean difference between focal and reference groups on the target ability

J_F = The sample size for the focal group

J_R = The sample size for the reference group

k_β = the degree of potential for bias (the mean difference between the two groups on the nuisance ability)

$\hat{\beta}_U$ = the estimated DTF/Test Bias

Table 8

Summary of Dimensionality and DTF/Test Bias Analyses

$c = 0.2$, $d_T = 0$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

Test Length	Dimensionality Analyses					DTF/Test Bias Analyses				
	T	p	Items Removed	T	p	Valid Subtest Length	Studied Subtest Length	$\hat{\beta}_U$	B	p
30	1.1	< .14	28-30	−1.6	< .95	27	3	0.27	6.9	< 00
	4.6	< 00	26-30	0.2	< .41	25	5	0.36	6.2	< 00
	5.5	< 00	25-30	0.5	< .30	24	6	0.46	6.8	< 00
40	3.8	< 00	37-40	−0.0	< .50	36	4	0.26	5.7	< 00
	7.0	< 00	35-40	−1.9	< .97	34	6	0.42	7.0	< 00
	7.7	< 00	33-40	−0.3	< .63	32	8	0.64	7.6	< 00
50	8.3	< 00	46-50	−0.2	< .58	45	5	0.59	9.1	< 00
	10.0	< 00	43-50	0.2	< .41	42	8	0.84	8.5	< 00
	9.7	< 00	41-50	1.8	< .04	40	10	0.86	7 9	< 00

Table 9

Summary of Dimensionality and DTF/Test Bias Analyses

$c = 0$, $d_T = 0.5$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

	Dimensionality Analyses					DTF/Test Bias Analyses				
Test Length	T	p	Items Removed	T	p	Valid Subtest Length	Studied Subtest Length	$\hat{\beta}_U$	B	p
30	0.9	< .18	28-30	−0.7	< .75	27	3	0.23	6.4	< 00
	4.7	< 00	26-30	−0.2	< .59	25	5	0.32	6.4	< 00
	5.2	< 00	25-30	−0.4	< .64	24	6	0.43	7.3	< 00
40	3.7	< 00	37-40	−1.2	< .89	36	4	0.33	7.5	< 00
	5.4	< 00	35-40	0.4	< .33	34	6	0.50	8.2	< 00
	7.5	< 00	33-40	−0.9	< .82	32	8	0.72	9.7	< 00
50	8.6	< 00	46-50	−0.5	< .67	45	5	0.41	6.8	< 00
	9.7	< 00	43-50	−0.3	< .63	42	8	0.63	7.3	< 00
	9.5	< 00	41-50	−1.2	< .89	40	10	0.76	7.8	< 00

Table 10

Summary of Dimensionality and DTF/Test Bias Analyses

$c = 0.2$, $d_T = 0.5$, $J_F = 1000$, $J_R = 1000$, $k_\beta = 0.5$

	Dimensionality Analyses					DTF/Test Bias Analyses				
Test Length	T	p	Items Removed	T	p	Valid Subtest Length	Studied Subtest Length	$\hat{\beta}_U$	B	p
30	1.9	< .00	28-30	−0.4	< .64	27	3	0.27	6.7	< 00
	4.6	< 00	26-30	−0.2	< .59	25	5	0.32	5.6	< 00
	5.3	< 00	25-30	1.1	< .14	24	6	0.50	7.7	< 00
40	4.0	< 00	37-40	−0.9	< .82	36	4	0.17	3.7	< 00
	6.4	< 00	35-40	−0.0	< .51	34	6	0.54	8.9	< 00
	8.1	< 00	33-40	−0.9	< .81	32	8	0.80	9.1	< 00
50	8.1	< 00	46-50	−1.3	< .09	45	5	0.57	8.3	< 00
	10.1	< 00	43-50	−0.4	< .67	42	8	0.80	8.5	< 00
	9.5	< 00	41-50	−1.2	< .89	40	10	0.97	8.5	< 00