A Multidimensionality-Based DIF Analysis Paradigm

Louis Roussos, Law School Admission Council
William Stout, University of Illinois at Urbana-Champaign

A multidimensionality-based differential item functioning (DIF) analysis paradigm is presented that unifies the substantive and statistical DIF analysis approaches by linking both to a theoretically sound and mathematically rigorous multidimensional conceptualization of DIF. This paradigm has the potential (1) to improve understanding of the causes of DIF by formulating and testing substantive dimensionality-based DIF hypotheses; (2) to reduce Type 1 error through a better understanding of the possible multidimensionality of an appropriate matching criterion; and (3) to increase power through the testing of bundles of items measuring similar dimensions. Using this approach, DIF analysis is shown to have the potential for greater integration in the overall test development process. *Index terms: bias, bundle DIF, cluster analysis, DIF estimation, DIF hypothesis testing, differential item functioning, dimensionality, DIMTEST, item response theory, multidimensionality, sensitivity review, SIBTEST.*

Differential item functioning (DIF) occurs in an ~~item~~ when examinees of equal trait levels (on the construct, or constructs, ~~the test~~ is intended to measure) but from separate populations differ in their probability of answering the item correctly. Many standardized ability or achievement tests (such as logical reasoning, reading comprehension, sentence completion, analytical reasoning, or mathematics word problems) require examinees to apply the ability that is being measured in context rather than abstractly or symbolically. This requirement provides diversity that enriches these tests and enhances their validity, because these tests involve measuring skills or abilities that must be applied in meaningful contexts. Thus, to demonstrate mastery, examinees should be required to demonstrate their ability to apply their skills in a variety of contexts. However, although it is desirable to embed the items of such tests in real-world situations, the presence of these contexts poses potential problems with respect to DIF.

Thus, one of the tasks of psychometricians and item writers is to ensure that the items are not posed within a domain that will offer a sizable advantage to particular subgroups of the test-taking population. This task requires careful dissection of the item statements and consideration of how familiarity with the context of the problem can be incorporated into the strategies of solving the problem, and whether group differences are present in knowledge within these contextual domains. To ensure that such tests are free of DIF, two distinct approaches to DIF analysis have arisen—one concerned with the development of statistical procedures for the detection of DIF items and one concerned with the development of substantive procedures for the design of DIF-free items.

The substantive DIF analysis approach has been constrained by the lack of statistical confirmation of its hypotheses about DIF-causing dimensions and by the lack of the development of new substantive hypotheses that can arise from exploratory statistical DIF analysis. The statistical DIF analysis approach has suffered from both lack of power caused by focusing mainly on the exploratory analysis of single items and Type 1 error inflation (inflated false alarm rate) caused by the use of inappropriate criteria for matching examinees. Much less attention has been given to exploratory and confirmatory statistical analyses of item bundles flowing from substantive dimensionality considerations. A *bundle* is any set of items selected according to some

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 4, December 1996, pp. 355–371

0146-6216/96/040355-17$2.10

organizing principle. These items are not necessarily adjacent nor do they necessarily refer to a common passage or text.

Multidimensionality-Based DIF

A new multidimensionality-based DIF analysis paradigm is presented here that unifies the substantive and statistical DIF analysis approaches by linking both to a theoretically sound and mathematically rigorous multidimensional conceptualization of DIF. With this paradigm, DIF analysis is shown to have the potential for greater participation in and improvement of the overall test development process.

The paradigm consists of a two-stage process. The first stage is a substantive DIF analysis, consisting of the development of DIF hypotheses based on the Shealy-Stout multidimensional model for DIF (MMD; Shealy & Stout, 1993a). "Model for DIF" means some way of linking substantive item characteristics (such as descriptive content, solution strategies required, superficial features, cognitive processes used, and so forth) to whether or to how much DIF will be displayed by an item for particular subgroups of the test-taking population. "DIF hypothesis" means a hypothesis of whether a particular combination of substantive item characteristics will give rise to DIF in an item or in a bundle of items; the term "DIF hypothesis" will be more rigorously defined below. These substantive item characteristics enter into this first stage of the DIF analysis by the initial stages of the test development process. In particular, these characteristics may arise from the test specifications or they may arise from test specialists listing item characteristics that are seen as potentially DIF producing.

The second stage of the DIF analysis consists of confirmatory statistical testing of the DIF hypotheses that resulted from the substantive DIF analysis in the first stage. The results of the second stage then feed back into the test development process for possible refinement of the test specifications and to decide what other substantive item characteristics should be allowed or excluded on the test. Thus, through the use of DIF hypotheses based on the MMD, the substantive and statistical DIF analyses are unified to form a single DIF analysis unit that can play a more integrated and helpful role in the test development process than is usually attributed to DIF.

DIF Terminology

The terminology established in Stout & Roussos (1995) will be used here. The term *dimension* will be used here to refer to any substantive characteristic of an item that can affect the probability of a correct response on the item (see Shealy & Stout, 1993a, for a more formal definition of dimension). The main construct(s) that the test is intended to measure (i.e., those usually scored) will be referred to as the *primary dimension(s)* of the test. It has long been recognized and accepted that the general cause of DIF is the presence of multidimensionality in items displaying DIF; that is, such items measure at least one dimension in addition to the primary dimension(s) the item is intended to measure (e.g., Berk, 1982; Cronbach, 1990; Dorans & Schmitt, 1989; Jensen, 1980; Lord, 1980; Messick, 1989; Scheuneman, 1982; Shepard, 1987; Wiley, 1990).

The additional possibly DIF-inducing dimensions will be referred to as *secondary dimensions*. Each secondary dimension is further categorized as either an *auxiliary* dimension if the secondary dimension is intended to be measured (perhaps as mandated by test specifications) or as a *nuisance* dimension if the secondary dimension is not intended to be measured; for example, the context of a logical reasoning item in a test that does not include context in the test specifications.

DIF that is caused by an auxiliary dimension will be referred to as *benign DIF*; DIF caused by a nuisance dimension will be referred to as *adverse DIF*. DIF caused by an auxiliary dimension is considered benign because the test is intended to measure the auxiliary dimension; however, as pointed out by Linn (1993), "the burden should be on those who want to retain an item with high DIF to provide a justification in terms

of the intended purposes of the test" (p. 353). Moreover, benign DIF is not necessarily ignorable because auxiliary dimensions with large benign DIF can sometimes be replaced by equally valid auxiliary dimensions with less DIF, or the number of items per auxiliary dimension can be modified to reduce the overall amount of benign DIF. DIF caused by a nuisance dimension is considered adverse because the item is less valid for one demographic population of examinees than for another in assessing examinee differences on the dimension(s) that is intended to be measured.

The demographic populations of interest in DIF analyses are usually either ethnicity-based or gender-based or a combination of both. For DIF analysis purposes, the populations are usually further categorized into a reference group population and a set of focal group populations. The term *focal* refers to the particular group of interest for the DIF analysis (e.g., African Americans, Asian Americans, females), and *reference* refers to the group with whom the focal group is to be compared (e.g., whites, males).

An Informal Multidimensional Model for DIF: DIF Resulting From Offensive Contextual Material or Favored Dimensions

A model for DIF is a procedure for linking the substantive characteristics of an item to its level of manifest DIF for particular reference and focal groups in the test-taking population. As indicated above, the multidimensional nature of DIF has long been recognized in the literature; however, formal mathematical models for DIF, such as Shealy & Stout (1993a), have only more recently been developed. Thus, it is not surprising that the predominant model for DIF that has been evident in attempts to explain the underlying causes of statistically detected DIF has been a less formal multidimensional conceptualization, but the model behind the explanations has never been explicitly stated. In order to more clearly compare this informal model for DIF with the model of Shealy and Stout, this informal model will be explained more explicitly here. Recognizing the multidimensional nature of DIF, this informal model has consisted of two components:

I1. The item measures a secondary dimension in addition to the primary dimension of the test that the item is intended to measure.

I2. The secondary dimension (when considered in isolation from the primary dimension) either advantages or disadvantages either the reference group or the focal group.

In regard to Component I2, a secondary dimension may advantage the reference or focal group when an item contains, for example, a content area for which one of the two groups may, on average, be more knowledgeable or familiar. For example, a logical reasoning item having a football context would be judged to favor males over females because, on average, males are more familiar with and knowledgeable about this content domain.

In the case of disadvantaging one of the two groups, the secondary dimension may be a content area that elicits an inappropriate emotional response that distracts the examinee, preventing or clouding the cognitive processes needed to solve the item. The March 3, 1995, issue of *The Chronicle of Higher Education* provided an extreme example of such an item. According to *The Chronicle*, a physics professor at MIT recently included the following question on an exam:

> You are in the forefront of a civil rights demonstration when the police decide to disperse the crowd using a water cannon. If the cannon they turn on you delivers 1000 liters of water per minute, what force does the water exert on you? (p. A-33)

As reported in *The Chronicle,* a senior in the Black Student Union at MIT pointed out, "If a question like that draws out an emotional response in a student, his or her mind is no longer on Physics" (p. A-33).

If Conditions I1 and I2 hold, the informal multidimensional model for DIF predicts that the item will manifest DIF. In fact, however, as will be discussed below, the MMD shows that somewhat paradoxically it is possible for both I1 and I2 to hold true in a DIF-free item, and it is possible for an item to display DIF even though I1 holds true but I2 does not.

Shealy and Stout's Multidimensional Model for DIF

Shealy & Stout (1993a) presented a MMD (similar to the model of Kok, 1988) that provides a rigorous mathematical definition of how latent trait parameters can cause DIF to become manifest in observed item responses. Shealy and Stout argued that DIF is due to the presence of two factors:

SS1. The item is sensitive not only to the construct θ that the item is intended to measure, but also to some secondary construct η, which may, in general, be multidimensional.

SS2. A difference exists between the two demographic groups of interest in their conditional distributions on η given a fixed value of θ (i.e., $\eta|\theta$).

If both of these factors are present, then DIF will likely occur; otherwise no DIF will occur. Table 1 summarizes the effect of the presence and absence of these two influences on DIF (at θ) for an item. Here $P(\theta)$ indicates that the probability of a correct response depends only on θ and not on η. There can, in principle, be a different amount of DIF at different values of θ. For example, this occurs in crossing DIF in which one group is favored at lower values and the other group is favored at higher values. The precise details of how the primary and secondary dimensions can interact to yield crossing DIF are discussed in more detail below; a more complete discussion can be found in Li & Stout (1996).

Table 1
Effect on DIF of Two Factors From the Shealy-Stout MMD

Conditional Probability Density Functions	Probability Correct	
	$P(\theta,\eta)$	$P(\theta)$
$f_R(\eta\|\theta) \neq f_F(\eta\|\theta)$	DIF	No DIF
$f_R(\eta\|\theta) = f_F(\eta\|\theta)$	No DIF	No DIF

Thus, the MMD hinges on two functions: (1) the function relating the probability of a correct response to the traits θ and η, denoted by $P(\theta,\eta)$, and (2) the conditional probability density function of η for examinees with fixed θ on the primary dimension, denoted by $f_G(\eta|\theta)$ (the subscript G denotes a subpopulation of the entire population of examinees; G = F indicates the focal group and G = R indicates the reference group). The function $P(\theta,\eta)$ does not depend on group membership because once all relevant examinee trait dimensions for answering the item are known, group membership is, by definition of the completeness of the latent trait space, no longer a predictor of performance on the item.

Thus, an item will exhibit DIF only if the reference and focal groups that have been equated on the primary dimension differ in distribution on a secondary dimension and if the item response function (IRF) for the item is sensitive to such a secondary dimension. Detailed examples are provided below.

If an item is sensitive to a secondary dimension, but the conditional distributions of the reference and focal groups do not differ on this secondary dimension, no manifest DIF will occur. If the examinees differ in their conditional distributions on a secondary dimension but the item is not sensitive to the secondary dimension (even though the secondary dimension may be present in the item), then no manifest DIF will occur. An example of the situation in which a secondary dimension is present in an item but the item response probability is not sensitive to it can be seen in the following logical reasoning item stimulus: "If that insect is a bee, it can only sting once. It only did sting once. So it is a bee." For this item, examinees are asked to select among multiple choices to find a syllogism with a similar pattern of reasoning. Although an obvious secondary dimension here is entomology, or knowledge of bees in particular, the wording of the item is such that it is difficult to imagine how any special knowledge of bees or entomology could possibly benefit examinees in solving this item. Thus, for this item it would seem reasonable to assume that $P(\theta,\eta)$ reduces to $P(\theta)$.

Condition SS1 is actually the same as I1 in the informal model; however, SS2 is different from I2. I2 refers to a difference in the η distributions for the two groups; SS2 refers to a difference in the $\eta|\theta$ distribu-

tions. As will be discussed in more detail below, it is possible for two groups to have the same η distributions (a secondary dimension that by itself does not favor either group) but to have the corresponding $\eta|\theta$ distributions different for the two groups. Similarly, it is possible for two groups to have different η distributions but the corresponding $\eta|\theta$ distributions are the same.

As will be shown below, this mathematical model fosters new insights into the design of DIF-free items in substantive DIF analysis and new insights into the statistical detection of DIF. For didactic purposes, each potentially DIF item will be assumed to be influenced by at most two dimensions—a unidimensional primary dimension θ and a single secondary dimension η. That is, (θ, η) is the complete latent space for the item.

Substantive DIF Analysis

Using the above terminology, "DIF hypothesis" is now defined more precisely as a hypothesis of whether an item (or bundle of items) that is sensitive to a particular primary dimension in combination with a particular secondary dimension will exhibit DIF for a particular pair of reference and focal groups and, if so, which group will be favored.

Current Methods

To better understand the current methods used in the development of DIF-free items, it is helpful to review how, in general, tests are constructed.

Test Specifications

An example of the dominant model for standardized test construction is the procedure followed by the major testing companies involved in measuring educational achievement [e.g., Educational Testing Service (ETS), American College Testing (ACT), and California Test Bureau (CTB)], which is based on what Messick (1989) referred to as curriculum relevance and representativeness. Thus, a set of test specifications is developed that "...reflects the *generality* of extant curricula in the subject-matter domain" to be tested (Messick, 1989, p. 65, emphasis as in the original text).

For example, consider the ACT Mathematics Assessment test (American College Testing, 1996). Item writing for this test is based on a two-way matrix of test specifications: content area × skill area. The content areas are pre-algebra, elementary algebra, intermediate algebra, coordinate geometry, plane geometry, and trigonometry. The items in each content area are further divided into three different skill areas: basic skills, application skills, and analysis skills. The test specifications delineate how many items are to be developed for each cell of the test specification matrix.

Similar to achievement tests, other standardized tests also involve a list of test specifications. An example of test specifications for a standardized verbal reasoning test are those for the verbal exam of the Scholastic Assessment Test (SATV) (College Board, 1995). The SATV consists of four types of items: reading comprehension, sentence completion, analogy, and antonym. The reading comprehension items are divided into six content categories (one of which is required to have a minority group orientation): narrative, biological science, argumentative, humanities, social science, and physical science. The other three types of SATV items are divided into four content areas: aesthetics/philosophy, practical affairs, science, and human relationships. The content area of aesthetics/philosophy includes music, art, architecture, literature, and philosophy. The content area of practical affairs includes money, tools, mechanical objects, sports, and historical topics. The actual writing of the items is performed by content experts (such as high school math teachers) following the substantive test specifications.

The above two examples demonstrate that standardized achievement or reasoning tests are typically designed for a single primary (often quite broad) dimension such as mathematics proficiency or verbal reasoning proficiency. Also, such tests typically contain a number of secondary dimensions as represented

by the content areas in the above two examples. These secondary dimensions appear to be auxiliary dimensions in the sense that their presence in the test specifications clearly indicates that they have an intentional impact on the measurement of the primary dimension.

However, note that in examining the items in an ACT math or SATV exam, a number of dimensions that are not present in the test specifications could be identified—these are the dimensions that are the possible nuisance dimensions of a test. For a math test, the context of algebra word problems might not be included in the specifications. For a verbal test, a content area may be so broadly specified (such as science) that subareas (such as environmental science) sometimes appear that are in accordance with the broader specification but result in unexpected DIF.

Sensitivity Review

Attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with a few exceptions, met with overwhelming failure (Cole, 1981; Engelhard, Hansche, & Rutledge, 1990; Linn, 1986; Plake, 1980; Reynolds, 1982; Shepard, 1981, 1982; Skaggs & Lissitz, 1992; Tittle, 1982). Only limited progress has occurred (Douglas, Roussos, & Stout, 1996; O'Neill & McPeek, 1993; Scheuneman, 1987; Schmitt, 1988; Schmitt, Holland, & Dorans, 1993), but so far few general principles for guiding item writing have been developed. The failure of substantive analyses for interpreting statistically identified DIF items for the underlying causes of DIF has led to the development of primarily subjective rules regarding the design of DIF-free items.

Perhaps the most notable example of such a set of rules is the *ETS Sensitivity Review Process: An Overview* (Educational Testing Service, 1987). The goals of the ETS sensitivity review process are to "...encourage the use of materials that acknowledge the contributions of women and minority group members" and "...to eliminate materials that women or minority group members are likely to find offensive or patronizing" (Linn, 1993, p. 356). Only the second of these two goals is relevant to eliminating DIF items, and it is directed only at eliminating certain possible causes of DIF, such as content offensive to females or ethnic minorities. The ETS sensitivity review process involves six criteria (Ramsey, 1993), five of which deal with eliminating offensive items and one (referred to as "balance") that on a test using secondary dimensions promotes the balanced use of secondary dimensions that reflects the diversity of the test-taking population. Because all ETS item writers are trained in the sensitivity review process, the review process does have a direct effect on the design of items for ETS tests. Unfortunately, as noted by Ramsey (1993), the sensitivity review process at ETS does not have any formal interaction with the statistical DIF analysis process. Thus, the statistical confirming or disconfirming of substantive DIF hypotheses based on the theoretical conceptualization of DIF inherent in the sensitivity review process (such as offensiveness of passage content) does not occur at ETS—"...there is no consistent effort to inform sensitivity reviewers what we are learning from DIF" (Ramsey, 1993, p. 385).

Substantive DIF Analysis From the Multidimensional Perspective: The Formation of Multidimensionality-Based DIF Hypotheses and the Design of DIF-Free Items

The MMD is proposed as a new approach to the design of DIF-free items. It is based on the idea that such a design should be linked to a theoretically sound multidimensional conceptualization of DIF. Thus, the proposal calls for the design of DIF-free items through the development of DIF hypotheses based on the nature of the underlying dimensionality structure of the test. Because the DIF hypotheses come from the model, the focus of this section will be on how to develop DIF hypotheses based on the MMD.

Using the MMD to Understand the Effect of Secondary Dimensions on Designing DIF-Free Items

As noted above, each potentially DIF item is assumed to be influenced by at most two dimensions—a

unidimensional primary dimension θ and a single secondary dimension η, which can vary from item to item. Furthermore, the joint distribution of the trait vector (θ, η) is assumed to be bivariate normal, with means μ_θ and μ_η, standard deviations (SDs) σ_θ and σ_η, and correlation ρ. For the reference group (R) and focal (F) group populations of interest, all parameters of this bivariate distribution are allowed to differ.

Under these assumptions, the expected difference in the means of η for examinees with a fixed value of θ for the two groups can be written as follows (e.g., Johnson & Wichern, 1992):

$$E_R(\eta|\theta) - E_F(\eta|\theta) = \left(\mu_{\eta_R} - \mu_{\eta_F}\right) + \theta\left(\rho_R \frac{\sigma_{\eta_R}}{\sigma_{\theta_R}} - \rho_F \frac{\sigma_{\eta_F}}{\sigma_{\theta_F}}\right) + \left(\mu_{\theta_F}\rho_F \frac{\sigma_{\eta_F}}{\sigma_{\theta_F}} - \mu_{\theta_R}\rho_R \frac{\sigma_{\eta_R}}{\sigma_{\theta_R}}\right). \quad (1)$$

In the MMD, DIF manifests itself through differences in the marginalized IRFs, $P_R(\theta)$ and $P_F(\theta)$:

$$P_R(\theta) = \int P(\theta,\eta) f_R(\eta|\theta) d\eta, \quad (2)$$

and

$$P_F(\theta) = \int P(\theta,\eta) f_F(\eta|\theta) d\eta. \quad (3)$$

The probability of correctly answering an item as a function of θ is thus obtained by averaging the response function $P(\theta, \eta)$ over the distribution of η for each fixed value of θ. The level of DIF at each θ is entirely accounted for by differences in the conditional distribution of $\eta|\theta$ for the reference and focal groups. Clearly, for fixed θ, $P_R(\theta)$ may not equal $P_F(\theta)$ when the conditional distributions $f_R(\eta|\theta) \neq f_F(\eta|\theta)$. Examining the bivariate normal model for the joint distribution of η and θ, it is often helpful to consider the special case in which both the reference and focal groups have the same SDs ($\sigma_{\theta_R} = \sigma_{\theta_F}$ and $\sigma_{\eta_R} = \sigma_{\eta_F}$) and correlation $\rho(\rho_R = \rho_F)$ and where $\sigma_\eta = \sigma_\theta$ (this is a totally nonrestrictive assumption). Hence, the two groups differ only in their means, μ_η and μ_θ. In this special case, Equation 1 reduces to:

$$E_R(\eta|\theta) - E_F(\eta|\theta) = \left(\mu_{\eta_R} - \mu_{\eta_F}\right) - \rho\left(\mu_{\theta_R} - \mu_{\theta_F}\right). \quad (4)$$

The difference in the conditional means given by Equation 4 contains all of the relevant information resulting from the difference between the conditional densities f_R and f_F. In fact, using a probabilistic principle known as stochastic ordering (Lehmann, 1959), it can be shown under the model of Equation 4 that the population G with the larger conditional mean for $\eta|\theta$ will have a larger value of $P_G(\theta)$. In most of the following examples, the focus is on the difference in mean values of $\eta|\theta$ for the reference and focal groups to illustrate how secondary dimensions can lead to DIF. For the special case represented by Equation 4, this difference is just the difference in the unconditional means of η minus the difference in the θ means. The latter difference is weighted by the correlation coefficient of the two latent dimensions. It is almost always reasonable to think of cognitive abilities as positively correlated ($\rho > 0$), which plays an important role in the analysis that follows.

For many achievement or reasoning tests, proficiency on the primary dimension (θ) is interpreted as the latent performance level on items that involve reasoning in context, such as reading comprehension. Depending on the test and the test-taking populations of interest, the test scores for the reference and focal groups may exhibit mean differences ranging from as much as 1 SD to values close to 0—sometimes the difference indicates higher reference group proficiency and sometimes it indicates higher focal group proficiency.

For example, scores on the SATV (College Board, 1995) suggest that for this particular test-taking population, females have a slightly higher primary dimension trait distribution than males ($\mu_{\theta_F} > \mu_{\theta_R}$). However, for the quantitative section of the SAT (SATQ) (College Board, 1995), males seem to have a slightly higher primary dimension trait distribution ($\mu_{\theta_R} > \mu_{\theta_F}$). These differences could be due to many factors, such as

different high school educational backgrounds between college-bound males and females.

Five Types of DIF Results and Their Interpretations Based on Item Wording

Case A. An item is observed to display unidirectional DIF in agreement with the informal method (discussed above) of predicting DIF, based on the presence of a secondary dimension (as suggested by the item wording) that favors either the reference or focal group.

Case B. An item displays unidirectional DIF in spite of the fact that the informal method of predicting DIF indicates there should be no DIF because the secondary dimension present clearly does not favor either the reference or focal group.

Case C. An item displays no DIF in agreement with the informal method of predicting DIF, based on either the absence of a secondary dimension influencing the item or the presence of a secondary dimension that favors neither the reference nor focal group.

Case D. An item does not display DIF, in spite of the fact that the informal method of predicting DIF indicates that DIF should be present because of the presence of a secondary dimension that clearly favors one of the groups.

Case E. An item displays crossing DIF, in which one group is favored in the low ability range and the other group in the high ability range. The informal method (which fails to view DIF as local in θ) provides no substantive explanations of crossing DIF.

Possible Causes of DIF

Case A. Because this is a case of manifest DIF, the Case A DIF items must be sensitive to a secondary dimension and the examinees must differ in their conditional distributions on the secondary dimension $[f_R(\eta|\theta) \neq f_F(\eta|\theta)]$. According to O'Neill & McPeek (1993), SAT reading comprehension (θ) items corresponding to reading passages that have content related to technical aspects of science (η) are an example of Case A items when the reference group is males and the focal group is females.

The informal dimensionality-based approach would suggest that DIF occurs because $\mu_{\eta_R} \neq \mu_{\eta_F}$. In this example, because college-bound males, on average, take more science courses than college-bound females, the informal approach assumes the DIF is due to $\mu_{\eta_R} > \mu_{\eta_F}$.

For the MMD, when $\mu_{\eta_R} \neq \mu_{\eta_F}$, DIF is indeed likely to occur when $\rho(\mu_{\theta_R} - \mu_{\theta_F})$ is not of the same sign and magnitude as $\mu_{\eta_R} - \mu_{\eta_F}$. In this example, note that females generally perform slightly better than males on the SATV, implying that $\rho(\mu_{\theta_R} - \mu_{\theta_F})$, although small, is of the opposite sign from $\mu_{\eta_R} - \mu_{\eta_F}$, which increases the potential for DIF against females.

Hence, the MMD results in the valuable added insight that a secondary dimension that causes DIF with respect to one primary dimension may not necessarily cause DIF with respect to another primary dimension; the MMD also adds some practical understanding as to why this can occur. Specifically, the MMD indicates that this secondary dimension will be less likely to cause manifest DIF when $\mu_{\theta_R} - \mu_{\theta_F}$ is of the same sign as $\mu_{\eta_R} - \mu_{\eta_F}$, and will be more likely to do so when they are of opposite signs. With respect to the formulation of DIF hypotheses, this discussion reveals that the informal method of merely assessing the secondary dimension regardless of the primary dimension is overly simplistic. The MMD clearly indicates that in forming DIF hypotheses, the sign of the difference in means on the primary dimension and on the secondary dimension, and the magnitude of the correlation between the secondary and primary dimensions, must be considered.

Case B. As in Case A, Case B is a case of manifest DIF in which the items are sensitive to a secondary dimension and the examinees differ in their conditional distributions on the secondary dimension. In Case B, an analysis by the informal method of DIF interpretation reveals that $\mu_{\eta_R} \approx \mu_{\eta_F}$. Thus, in using the informal method it is concluded that the secondary dimension should not induce manifest DIF and the item

becomes one of those systemic items for which no explanation of the manifest DIF can be found. One possible example of this is the reported DIF (O'Neill & McPeek, 1993) in favor of African-Americans relative to whites on analogy and antonym items on the SAT and the Graduate Record Exam (GRE) that deal with human relationships.

There is no clear argument as to why African-Americans would score higher than whites, or to why whites would score higher than African-Americans, in terms of interest or knowledge on the secondary dimension of "human relationships," which suggests that $\mu_{\eta_R} \approx \mu_{\eta_F}$ is likely. Thus, in this example the informal DIF interpretation method does not seem to produce any understanding as to why these items would exhibit the reported DIF.

However, referring to Equation 4, the MMD shows (paradoxically from the informal viewpoint) that even if $\mu_{\eta_R} \approx \mu_{\eta_F}$ the mere presence of a secondary dimension can induce DIF through its correlation with the primary dimension, if a difference in mean proficiency on the primary dimension exists between the two groups. In this example, because African-Americans, as a group, have lower average scores than whites on the SAT and GRE verbal sections, if $\mu_{\eta_R} \approx \mu_{\eta_F}$ and $\mu_{\theta_R} > \mu_{\theta_F}$, the MMD suggests DIF in favor of African-Americans. Thus, a common "unexplainable" DIF result could have a theoretically-based explanation when examined using the MMD. With respect to the formulation of DIF hypotheses, the MMD indicates that this type of DIF is not just explainable, it is actually predictable. With greater attention to how primary and secondary dimensions interact to form DIF as explained by the MMD, practitioners can hope to move from merely reactive DIF explanations (as in the above discussion) to proactive formulating and testing of DIF hypotheses.

Case C. Case C is a case of no manifest DIF. Hence, either the item is not sensitive to a secondary dimension (Case C1) or the item is sensitive to a secondary dimension but the two groups do not differ in their distributions on the secondary dimension once they have been equated on the primary dimension (Case C2).

Case C1 assumes that the item is not sensitive to a secondary dimension. The informal method predicts no manifest DIF because the item appears to be either measuring no secondary dimension or is insensitive to the secondary dimension that is present. Thus, the informal method correctly concludes that the IRF depends only on θ. Because the IRF depends only on θ, an inspection of the item using the MMD also correctly concludes that the item should display no manifest DIF. Thus, in this case the informal method of interpretation and the MMD are in complete agreement. An example of a test consisting of such items is a math test consisting of purely symbolic items measuring a narrowly defined area, such as factoring or the use of exponents.

For Case C2, suppose the item does depend on η but that an inspection of the item reveals that $\mu_{\eta_R} \approx \mu_{\eta_F}$. Then the informal method predicts no manifest DIF. Referring to the special case of Equation 4, the MMD shows that when $\mu_{\eta_R} \approx \mu_{\eta_F}$, no manifest DIF will occur as long as it is also true that either $\rho \approx 0$ (which seldom occurs in practice) or $\mu_{\theta_R} \approx \mu_{\theta_F}$. An example that comes close to this latter possibility is presented in Douglas et al. (1996) in which it was hypothesized with respect to male-female DIF that the secondary dimension "money and business" would slightly favor males. They then analyzed a bundle of eight items on a logical reasoning test on which males had only slightly higher scores than females. Thus, $\mu_{\theta_R} \approx \mu_{\theta_F}$ (scores for males were only slightly higher than for females) and $\mu_{\eta_R} \approx \mu_{\eta_F}$ (males were only slightly favored on the secondary dimension), which results in a prediction of little-to-no manifest DIF, which is what was found in their data analysis. If the slight advantage hypothesized for males (which was in the same direction for both θ and η) is taken into account, Equation 4 indicates that the two slight advantages would tend to cancel each other in terms of the expected difference in means on $\eta|\theta$, which still indicates that little-to-no manifest DIF should occur.

As mentioned in the discussion of Case B and reiterated here, the MMD also results in the added insight that a seemingly innocuous secondary dimension that results in no DIF with one primary dimension may

not necessarily result in no manifest DIF with other primary dimensions if it is positively correlated with other primary dimensions and if the reference and focal groups differ in trait levels on these other primary dimensions. Again, an advantage of the MMD is the emphasis on understanding how the secondary and primary dimensions interact rather than merely considering the secondary dimensions in isolation, which is seen as the key to not only explaining DIF but also predicting it.

Case D. Case D is a case of no manifest DIF. In Case D, the item is sensitive to a secondary dimension but the two groups do not differ in their distributions on the secondary dimension once they have been equated on the primary dimension, as the MMD requires. Case D is, perhaps, the most paradoxical of all the cases. Because the informal method of DIF interpretation correctly concludes that $\mu_{\eta_R} \neq \mu_{\eta_F}$, it supplies no explanation as to how the item could possibly result in no manifest DIF. The MMD, however, shows that if $\rho(\mu_{\theta_R} - \mu_{\theta_F})$ is of the same sign and of similar magnitude as $\mu_{\eta_R} - \mu_{\eta_F}$, then DIF will likely not become manifest for the item.

An example of this is provided by some seemingly contradictory findings of O'Neill & McPeek (1993) and Douglas et al. (1996). O'Neill and McPeek found that some types of items measuring verbal reasoning tend to exhibit DIF in favor of males if they concern "practical affairs," of which money was used as an example. Because males and females tend to perform approximately equally, or females perhaps slightly better, on the type of verbal exams that they considered (i.e., $\mu_{\theta_F} \geq \mu_{\theta_R}$), a finding of DIF in favor of males on these items would thus indicate by the informal method that $\mu_{\eta_R} > \mu_{\eta_F}$ must hold, where η denotes familiarity or knowledge of money and financial matters. By contrast, Douglas et al. (1996) reported that logical reasoning items within the context of money or finances show little DIF against either males or females.

As mentioned above, a close look at Equation 4 reveals a possible explanation for this apparent contradiction. For both the logical reasoning items and the verbal reasoning items, under the assumptions made here, the first term in Equation 4 dealing with the η means is positive. In the case of the verbal items in which females are more proficient on average, the second term (including the minus sign) is positive as well, almost ensuring DIF in favor of males. However, in the case of logical reasoning, in which males seem to have a slightly higher mean proficiency, the second term serves to cancel the difference created by the first term. The genuinely paradoxical, but logically correct, conclusion is that DIF can be reduced to some extent if the proficiency distribution on the secondary dimension tends to favor the same group that the proficiency distribution on the primary dimension does. Note, however, that this is a delicate matter, because a large difference in the η means (the first term in Equation 4) cannot be offset by the opposite difference in the θ means (the second term in Equation 4) especially if η and θ have a low correlation (ρ).

For example, the use of football word problems on a high school math test would not be justified. The large advantage of males over females on η (familiarity with and knowledge of football) would not come close to being cancelled by the much smaller advantage of males over females on θ (mathematics proficiency), even if the correlation between η and θ were large. Nonetheless, for other secondary dimensions, such as familiarity with physical science terminology, the model may prove a useful tool for selecting acceptable content areas to accompany a particular primary dimension to minimize the possible occurrence of large DIF.

As indicated by the MMD, the successful formulation of DIF hypotheses is contingent on knowing the dimensionality structure of the test. Test specifications provide a starting point. A construct validity analysis from Embretson's (1983) cognitive processing point of view is the ultimate goal. Expanding the test specifications to include more of the secondary dimensions of the test is an important first step. The above example also highlights that the MMD can explain why a secondary dimension ("money or finances" in the above example) can result in DIF on one test but not on another, which is another confusing finding often reported in the DIF literature.

Case E. As mentioned above, the informal method of DIF interpretation yields no explanation for

crossing DIF. However, by using Equation 1, an explanation for crossing DIF is readily obtained from the MMD. According to Equation 1, the expected difference in the means of the $\eta|\theta$ distributions for the two groups will be a function of θ on the primary dimension if

$$\rho_R \frac{\sigma_{\eta_R}}{\sigma_{\theta_R}} \neq \rho_F \frac{\sigma_{\eta_F}}{\sigma_{\theta_F}}. \quad (5)$$

Equation 5 provides insight as to the type of trait distributions that lead to crossing DIF for bivariate normal traits. For example, if $\rho_R = \rho_F$, one possible combination of parameters that could lead to crossing DIF would be $\sigma_{\theta_F} < \sigma_{\theta_R}$ combined with $\sigma_{\eta_F} > \sigma_{\eta_R}$. Thus, if the SD of the primary dimension is smaller in the first group than in the second and the SD of the secondary dimension is larger in the first group than in the second, there is a potential for crossing DIF. Of course, many other combinations of parameter values exist that can result in crossing DIF.

Construction of Substantive DIF Hypotheses Using the Partial Delineation of Dimensionality Distributional Structure of Primary and Secondary Dimensions

The conclusion from the above didactic analysis is that if the joint trait distributions of the reference and focal groups on the primary dimension and a secondary dimension can be estimated, then a DIF hypothesis can be formed concerning the secondary dimension, reasoning on the basis of Equation 4, for example. That is, if $E_R[\eta|\theta] - E_F[\eta|\theta]$ is positive, DIF against the focal group is hypothesized; if it is negative, DIF against the reference group is hypothesized. Note that because the DIF hypotheses can be merely directional (the magnitude of the DIF does not need to be specified), the level of detail at which the joint trait distribution information needs to be specified to develop useful DIF hypotheses can be quite low. The level of detail for estimating (1) the difference in secondary dimension means, (2) the difference in primary dimension means, and (3) the correlations between the primary and secondary dimensions, can range from merely estimating the signs of these quantities to estimating the approximate size of the quantities to estimating the quantities precisely. At any of these levels of detail, as long as some information is estimated about each of these three quantities, useful DIF hypotheses can be developed for the secondary dimensions. These hypotheses would then be tested and would result in a body of confirmed DIF hypotheses. Then, based on these confirmed DIF hypotheses, test developers would have the opportunity to consider modifying the test specifications by selecting alternative auxiliary secondary dimensions that still meet the specifications but display less DIF than auxiliary dimensions already being used. Also, test developers could prohibit the use of any nuisance secondary dimensions that tend to result in moderate to large amounts of DIF. There are four practical ways to obtain this multidimensional trait distribution information for use in formulating DIF hypotheses: (1) using previously published DIF analyses, (2) from purely theoretical substantive content considerations, (3) using archival test data, and (4) pretesting dimensions instead of just items.

Using previously published DIF analyses. DIF hypotheses for a variety of combinations of primary and secondary dimensions can be directly obtained from already published DIF analyses. For example, with respect to male-female DIF, O'Neill & McPeek (1993) reported that reading comprehension (θ) items tended to favor males when the content had to do with science (η_1) and to favor females when the content was related to the humanities (η_2).

From purely theoretical substantive content considerations. Based on substantive knowledge about certain secondary and/or primary dimensions, it is possible to hypothesize whether a particular group will be favored. For example, Douglas et al. (1996) evaluated the DIF of logical reasoning (θ) items that dealt with children (η). The test scores indicated that μ_{θ_R} was slightly greater than μ_{θ_F}. Based on the fact that females still bear the vast majority of child-rearing responsibility, μ_{η_F} would be theorized to be greater than μ_{η_R}. Assuming that θ and η are positively correlated leads to the prediction using Equation 4 that such items would

display DIF in favor of females, which was indeed the finding in their paper.

Using archival test data. For some secondary dimensions, standardized tests or subsets of standardized tests already exist that measure them. For example, the College Board Advanced Placement Tests include subject tests in the areas of art history and studio art, biology, chemistry, computer science, economics, English, French, German, government and politics, history, Latin, mathematics, music, physics, psychology, and Spanish. Determining marginal distributions for such secondary dimensions is readily within the grasp of organizations that construct such tests. The correlation of these secondary dimensions with any primary dimension can be assumed to be positive in the vast majority of cases. A particular correlation can often be directly estimated because such tests are frequently administered as a battery with a group of examinees being tested on a number of dimensions. If the archival data consist of a single test that contains auxiliary dimensions, joint trait level distributions for the primary and auxiliary secondary dimensions (means and correlations) can be obtained by subscoring the auxiliary dimensions. An example of such a test is the ACT mathematics test described above. In general, whenever tests have sufficient numbers of items measuring common secondary dimensions, obtaining primary-secondary joint distribution information is possible. Then, using these distributions, DIF hypotheses can be formulated based on Equation 1 or Equation 4.

Pretesting dimensions instead of just items. Pretest items are frequently included on standardized test administrations for determining the psychometric qualities of the pretest items in anticipation of using them on future forms. In the future, this same procedure of inserting nonoperational items could be used to test secondary dimensions of interest for DIF by investigating these dimensions in the form of item bundles (groups of items with common primary and secondary dimensions). The results of the exploratory statistical DIF analyses of the dimensions would lead to the formulation of dimensionality-based DIF hypotheses and to some information on the joint distribution of the primary and secondary dimensions. The accuracy of this information would depend on the size of the item bundles.

Using Confirmed DIF Hypotheses in Test Construction

Estimating the characteristics of the joint distributions of primary and secondary dimensions and developing the associated DIF hypotheses is one component required for implementing this approach for designing DIF-free items. The other component is coordinating the test construction process with the information resulting from the confirmed DIF hypotheses.

In general, from the multidimensional perspective of DIF, improved control over substantive item characteristics through improved test specifications is a key element in controlling DIF. In the case of auxiliary secondary dimensions, specifications might be improved by being made more specific or by considering alternative auxiliary dimensions. For example, instead of specifying "science" as a topic specification for a reading comprehension test, there is evidence from DIF analyses that differentiating between health-related/biological science and physical science is important (Douglas et al., 1996).

Some nuisance dimensions might be brought under stricter control by requiring their presence in the test specifications. For example, the varying contexts in word problems on an algebra test could be specified in the same way that the contexts of reading comprehension passages are specified. Nuisance dimensions that are not desired to be included in the usual "fixed" set of test specifications could be included in a second "variable" set. The variable specifications would consist of a list of acceptable nuisance dimensions from which any subset would be allowed on a test.

Statistical DIF Analysis

The new substantive DIF analysis approach to designing DIF-free items is inherently interactive with dimensionality-based statistical analyses for the estimation and detection of DIF. In particular, the above detailed discussion of how to use the MMD to develop DIF hypotheses is an integral part of the process of

statistically detecting DIF. Most importantly, it is the multidimensionality considerations underlying the DIF hypotheses that are the common link between the two analyses. These considerations have a number of implications for implementing statistical analyses to detect and estimate DIF.

Current Method: One-Item-at-a-Time DIF Analysis

The predominant method of DIF analysis for nationally administered standardized achievement tests consists of calculating a DIF statistic (most commonly the Mantel-Haenszel statistic as modified by Holland & Thayer, 1988) for each item on a test, and the matching criterion is total score on the test or on some subset of the items on the test. Often the matching criterion is "purified" of items suspected of DIF. Other than this purification of the matching criterion, dimensionality considerations generally play no further role in the selection of the matching criterion nor any role at all in the selection of the items to be tested for DIF.

DIF Estimation and Detection From the Multidimensional Perspective

As noted above, the design of standardized achievement tests involves test specifications which indicate that although the tests may be designed to have a single primary dimension, they also seem to be designed to have a number of secondary dimensions. As noted by O'Neill & McPeek (1993), "[i]f a test contains many different areas of knowledge or skill, examinees matched on the total test score are not necessarily matched on each separate area contained in the test."

O'Neill and McPeek point out that a high DIF value for an item can be due solely to "...the nature of the criterion used for matching" (p. 256). In other words, if the matching criterion is multidimensional, then a statistical test for DIF may reject a perfectly fair item simply because the examinees differ on one of the auxiliary dimensions (secondary dimensions intended to be measured by the test). Thus, with regard to Type 1 error, perhaps the most insidious cause has been the assumption of a unidimensional matching criterion when in fact the test is measuring either more than one primary dimension or a single primary dimension with several auxiliary secondary dimensions.

The classic example of this situation is the repeated finding on mathematics tests that geometry items exhibit DIF against females, with the opposite finding that algebra items exhibit DIF against males (see Doolittle & Cleary, 1987; O'Neill & McPeek, 1993). Clearly, this is a case in which the test is measuring either two primary dimensions (algebra and geometry) or a primary dimension of mathematics with two auxiliary dimensions (algebra and geometry); the DIF could disappear if examinees were matched separately on both dimensions instead of on the sum of the two dimensions as represented by the sum score on the entire test.

With regard to the above substantive DIF analysis discussion and to the above comment on Type 1 error, a growing body of research (Ackerman, 1992; Angoff, 1993; Linn, 1993; Mazor, Kanjee, & Clauser, 1995; Shealy & Stout, 1993a, 1993b; Skaggs & Lissitz, 1992) calls for DIF analyses that are predicated on a better understanding of the multiple cognitive dimensions that underlie tests, especially as espoused in the construct validity literature (e.g., Anastasi, 1986; Cronbach & Meehl, 1955; Embretson, 1983; Loevinger, 1957; Messick, 1989; Wiley, 1990). Hence, it is clearly very important to:

1. Consider the dimensionality structure of a test prior to conducting a DIF analysis.
2. Attempt to use as the matching criterion the set of items that most purely measures only the primary dimension(s) of the test or, failing that, use the items whose secondary dimensions are such that the items are most likely DIF-free, perhaps based on previously confirmed DIF hypotheses concerning items with particular secondary dimensions. If two scores are required for matching, then it is essential to use a DIF procedure such as SIBTEST (Stout, Li, & Nandakumar, 1996) or logistic regression (Mazor et al., 1995) that allows multiple score matching.
3. Distinguish between the secondary dimensions that the test is intended to measure (auxiliary dimensions) and the secondary dimensions that the test is not intended to measure (nuisance dimensions).

The dimensionality-based approach has two major advantages over the single-item exploratory DIF analyses that are currently conducted: (1) testing more than one item at a time results in greater statistical detection power when the items measure a common secondary dimension; and (2) the formation and testing of substantive DIF hypotheses naturally leads to a better understanding of the causes of DIF, to improved item writing and test design, and to a better understanding of differential knowledge structures across groups.

Dimensionality Analysis

Dimensionality analyses should be conducted that include (1) substantive dimensionality analyses for hypothesizing possible dimensionality structures and (2) statistical dimensionality analyses for confirming hypothesized dimensionality structures and for exploring new dimensionality hypotheses. These analyses should lead to the selection of items for the matching criteria that best represent the primary and/or auxiliary dimensions. Thus, it is conceivable that they could serve as matching criteria in a DIF analysis. The dimensionality analyses should also lead to candidate nuisance dimension item bundles for statistical DIF analysis. In spite of the long-standing recognition of the multidimensionality inherent in DIF, it is rare that a published DIF analysis includes a dimensionality analysis (Cohen & Kim, 1992; Douglas et al., 1996; Mazor et al., 1995), and even rarer for the dimensionality analysis to be used in the selection of the matching criterion and the items to be tested for DIF (Douglas et al., 1996).

DIF Analysis of the Dimensions Using Bundle DIF Analysis

Instead of analyzing items for DIF, it is proposed that dimensions be analyzed for DIF. This would be accomplished by analyzing item bundles—items that have been designed or selected to measure not only the same primary dimension but also the same secondary dimension. In the case of auxiliary dimensions, this could be accomplished quite readily by simply testing item bundles composed of items with a common test specification criterion. In the case of nuisance dimensions, there would need to be at least two items per nuisance dimension. For auxiliary dimensions, those with sufficient numbers of items (e.g., 20 or more) could be conditioned on for analyzing nuisance dimensions for DIF (e.g., conditioning on algebra and geometry items separately for a math test).

Two Examples of Dimensionality-Based DIF Analysis

In principle, the approach recommended here can be implemented with any DIF procedure that is capable of analyzing item bundles. Two such procedures are the parametric item response theory (IRT) DIF analysis procedure of Raju, van der Linden, & Fleer (1995) and the SIBTEST DIF procedure (Shealy & Stout, 1993b). For completeness, two examples of dimensionality-based DIF analysis methods are provided that were presented in Douglas et al. (1996) in which the SIBTEST DIF procedure was used.

Each of the two sample methods attempts to identify bundles with potential for DIF amplification by letting each bundle consist of items measuring a common secondary dimension in addition to the intended primary dimension. Each identified suspect bundle can then be statistically analyzed for bundle DIF. The first method involves a substantive subjective dimensionality analysis in which expert opinion (i.e., human judgment) is used to select dimensionally homogeneous bundles of items suspected of differential bundle functioning (DBF). The second method involves the use of a statistical dimensionality analysis to find the item bundles. In both methods, dimensionality considerations affect not only the selection of the items to be tested for DIF but also the selection of the items to be used as the matching criterion.

The first method presented in Douglas et al. (1996) of selecting item bundles that may have a common secondary dimension is to simply read the items and use expert opinion to group the items by content. As discussed above, DIF hypotheses based on expert opinion can arise from either empirical evidence accumulated over time or from content considerations developed from subjective evaluation, possibly using a

representative panel of expert judges. In Douglas et al. (1996), a nonexpert panel of judges was used because the example was intended only to illustrate the approach. Items judged as potentially DIF were lumped into bundles according to common secondary dimensions. The items judged as not potentially DIF became the proposed matching criterion. A standard one-item-at-a-time DIF analysis was also performed on this set of presumably DIF-free items. All items displaying negligible DIF were retained for use in the matching criterion. Finally, the SIBTEST procedure was run on the bundles to test the DIF hypotheses. [See also Oshima, Raju, Flowers, & Slinde (1996) for a similar example of a dimensionality-based DIF analysis of item bundles using the parametric IRT approach of Raju et al. (1995).]

In the second method used by Douglas et al. (1996), an exploratory statistical dimensionality analysis was used on one portion of the examinee data to develop DIF hypotheses, and then a confirmatory analysis on a cross-validation examinee portion was conducted using the expert opinion approach discussed above. The exploratory dimensionality analysis (using the methods of Stout, Douglas, Habing, Kim, Roussos, & Zhang, 1996) consisted of the combined use of hierarchical cluster analysis (HCA) and DIMTEST [the Stout and Nandakumar (Nandakumar & Stout, 1993; Stout, 1987) method of assessing the dimensional distinctiveness of two subtests of items]. This analysis resulted in item bundles corresponding to identified secondary dimensions. DIF hypotheses were formed by substantive inspection of the items in these bundles. The matching criterion for the DIF analysis was composed of the items that did not fall into these bundles. These DIF hypotheses were then subjected to a confirmatory statistical DIF analysis using SIBTEST. For a more detailed discussion of these two methods, see Douglas et al. (1996).

In summary, it is proposed that a clearer delineation of the dimensionality structure of a test be implemented at both the design and analysis stages of test development so as to result in DIF-free operational items (based on previous pretest results) and in the testing of suspected DIF-causing nuisance dimensions (using item bundles) at the pretest stage rather than merely suspected DIF items.

Summary: A New DIF Analysis Paradigm

A new DIF analysis paradigm is proposed that unifies the substantive and statistical DIF analysis approaches by linking both to a theoretically sound and mathematically rigorous multidimensional conceptualization of DIF. The paradigm consists of the development and testing of dimensionality-based DIF hypotheses using a combination of substantive and statistical analyses. The substantive analysis is used to develop DIF hypotheses, and the statistical analysis is used to test the DIF hypotheses and estimate the amount of DIF. This unified approach results in the potential for DIF analysis to be more closely integrated with the overall test development process. The natural unity of the substantive and statistical DIF analyses flows from the use of a formal multidimensional model for DIF, the MMD. The MMD has the potential (1) to improve the understanding of the causes of DIF through formulating and testing substantive dimensionality-based DIF hypotheses, (2) to reduce Type 1 error through a better understanding of the dimensionality of the matching criterion, and (3) to increase statistical detection power through the testing of bundles of items measuring similar primary and secondary dimensions. The approach also uses a test-development/DIF-analysis feedback loop in which the test development process feeds test specifications and lists of other substantive item characteristics into the DIF analysis, DIF hypotheses based on these characteristics are developed and tested, and the results are fed back into the test development process for possible modification of the item characteristics specified for or allowed on the test.

The interaction of substantive and statistical analyses at a testing organization could result in a real understanding of the causes of DIF. The proposed paradigm for the design of DIF-free items is a demanding, yet promising, proposal. One very simple step that could be taken immediately by any major testing organization is to ensure that a strong communication link exists between the DIF detection process and the sensitivity review process. Indeed, Ramsey (1993) has already recommended that such a step (establishing

a communication link between the statistical DIF analysis and the sensitivity review process) be taken at ETS. Linn (1993) wrote, "...it is not too early for ETS and its major clients to begin the process of opening up test specifications for a major review, taking into account not only what has been learned from DIF analysis but [also] what has been learned from the past several years of sensitivity reviews" (p. 364). All testing organizations could benefit from beginning to link their substantive and statistical DIF analyses.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

American College Testing. (1996). *Test preparation: Reference manual for teachers and counselors.* Iowa City IA: Author.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1–15.

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

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore MD: The Johns Hopkins University Press.

Cohen, A. S., & Kim, S.-H. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5,* 303–320.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

College Board. (1995). *Counselor's handbook for the SAT program.* New York: Author.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24,* 157–166.

Dorans, N. J., & Schmitt, A. P. (1989, March). *The methods for dimensionality assessment and DIF detection.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465–485.

Educational Testing Service. (1987). *ETS sensitivity review process: An overview.* Princeton NJ: Author.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347–360.

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

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

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs NJ: Prentice Hall.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–275). New York: Plenum.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

Li, H.-H., & Stout, W. F. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647–677.

Linn, R. L. (1986). Bias in college admissions. In *Measures in the college admissions process: A College Board colloquium* (pp. 80–86). New York: The College Entrance Examination Board.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale NJ: Erlbaum.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694 (Monograph Supplement, 9).

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

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131–144.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13–103). New York: Macmillan.

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

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–279). Hillsdale

NJ: Erlbaum.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1996, April). *Differential bundle functioning (DBF) using the DFIT framework: Procedures for identifying possible sources of DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Plake, B. S. (1980). A comparison of statistical and subjective procedures to ascertain item validity: One step in the validation process. *Educational and Psychological Measurement, 40,* 397–404.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale NJ: Erlbaum.

Reynolds, C. R. (1982). The problem of bias in psychological measurement. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–201). New York: Wiley.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 180–198). Baltimore MD: The Johns Hopkins University Press.

Scheuneman, J. D. (1987). An experimental exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97–118.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25,* 1–13.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale NJ: Erlbaum.

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

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

Shepard, L. A. (1981). Identifying bias in test items. In B. F. Green (Ed.), *New directions in testing and measurement: Issues in testing—Coaching, disclosure, and test bias* (No. 11; pp. 79–104). San Francisco: Jossey-Bass.

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

Shepard, L. A. (1987). Discussant comments on the NCME symposium: Unexpected differential item performance and its assessment among black, Asian-American, and Hispanic students. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton NJ: Educational Testing Service.

Skaggs, G., & Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29,* 227–242.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F., Douglas, J., Habing, B., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331–354.

Stout, W. F., Li., H.-H., & Nandakumar, R. (1996). *Use of SIBTEST to do DIF when the matching subtest is intentionally multidimensional.* Manuscript submitted for publication.

Stout, W. F., & Roussos, L. A. (1995). *SIBTEST users manual* (2nd ed.) [Computer program manual]. Urbana-Champaign: University of Illinois, Department of Statistics.

Tittle, C. K. (1982). Use of judgemental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 31–63). Baltimore MD: The Johns Hopkins University Press.

Wiley, D. E. (1990). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75–107). Hillsdale NJ: Erlbaum.

Acknowledgments

The authors thank the Special Issue editor and two anonymous reviewers for helpful comments, and especially acknowledge helpful discussions with Anne Seraphine of the University of Texas at Austin and Test Specialists Deborah Kerman, Stephen Luebke, and Theresa Robinson of the Law School Admission Council.

Author's Address

Send requests for reprints or further information to Louis Roussos, Law School Admission Council, 661 Penn Street, Newtown PA 18940, U.S.A. or to William Stout, Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois, Champaign IL 61820, U.S.A. Email: lroussos@lsac.org or stout@stat.uiuc.edu.