

Penfield, R., Alvarez, K., Lee, O. (in press). Using a taxonomy of differential step functioning form to improve the interpretation of DIF in polytomous items. *Applied Measurement in Education*.

Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF  
in Polytomous Items: An Illustration

## **Abstract**

The assessment of differential item functioning (DIF) in polytomous items addresses between-group differences in measurement properties at the item level, but typically does not inform which score levels may be involved in the DIF effect. The framework of differential step functioning (DSF) addresses this issue by examining between-group differences in the measurement properties at each step underlying the polytomous response variable. The pattern of the DSF effects across the steps of the polytomous response variable can assume several different forms, and the different forms can have different implications for the sensitivity of DIF detection and the final interpretation of the causes of the DIF effect. In this paper we propose a taxonomy of DSF forms, establish guidelines for using the form of DSF to help target and guide item content review and item revision, and provide procedural rules for using the frameworks of DSF and DIF in tandem to yield a comprehensive assessment of between-group measurement equivalence in polytomous items.

## **Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF in Polytomous Items: An Illustration**

The framework of differential item functioning (DIF) has become an integral component of test validation methodology and the study of test fairness (AERA/APA/NCME, 1999). The presence of DIF in a particular item indicates that individuals having the same level of ability, but belonging to different groups, do not share the same expected response to the item (Holland & Thayer, 1988; Penfield & Camilli, 2007; Roussos & Stout, 2004). By convention, the item under investigation for DIF is called the studied item, and the two groups being compared are called the *reference* and *focal* groups (historically, the reference group is the group for which the test is expected to favor and focal group is the targeted or disadvantaged group of interest). Numerous procedures for studying DIF in dichotomous items have been proposed (Camilli & Shepard, 1994, Clauser & Mazor, 1998; Penfield & Camilli, 2007), and the application of DIF approaches to dichotomously scored items is well documented in the literature (Camilli & Shepard, 1994; Holland & Wainer, 1993; Penfield & Camilli, 2007; Roussos & Stout, 2004).

Although incipient procedures for assessing DIF focused on dichotomous items, the increasing use of polytomous item formats (i.e., performance based assessments, innovative item types, etc.) has led to the development of numerous methods for assessing DIF in polytomous items (Penfield & Camilli, 2007; Penfield & Lam, 2000; Potenza & Dorans, 1995). Methods for assessing DIF in polytomous items focus on whether a between-group difference in the measurement properties exists somewhere within the polytomous response variable, but typically provide no information as to

precisely which score levels are implicated in this difference. For this reason, traditional DIF indices for polytomous items can be conceptualized as omnibus indices of measurement equivalence.

Two general classes of omnibus polytomous DIF statistics exist: *net* DIF statistics and *global* DIF statistics. Net polytomous DIF statistics measure and assess the aggregated net effect across all score levels of the polytomous items, and thus a difference in sign of the DIF effect across the score levels (e.g., DIF favoring the reference group for some score levels but favoring the focal group for others) can yield a net DIF effect of zero (or near zero) despite the presence of sizable effects within particular score levels. Examples of net polytomous DIF statistics include the standardized mean difference index (Dorans & Schmitt, 1993), the polytomous SIBTEST approach (Chang, Mazzeo, & Roussos, 1996), the polytomous DFIT approach (Flowers, Oshima, & Raju, 1999), Mantel's chi-square test (Mantel, 1963; Zwick, Donoghue, & Grima, 1993), and the cumulative common log-odds ratio (Liu & Agresti, 1996; Penfield & Algina, 2003). In contrast, global DIF statistics assess the unsigned effect across all score levels of the polytomous item. Examples of global polytomous DIF statistics include the generalized Mantel-Haenszel statistic (GMH; Somes, 1986; Zwick et al., 1993), the polytomous logistic regression approach (French & Miller, 1996), item response theory (IRT) likelihood ratio tests (Ankenmann, Witt, & Dunbar, 1999; Kim & Cohen, 1998), and the simultaneous step-level (SSL) test of DIF (Penfield, in press). It is germane to note that while global DIF approaches are sensitive to DIF effects that change sign across the score levels, they still provide no information concerning precisely which score levels are responsible for the DIF effect observed in the item.

Given that omnibus measures of DIF for polytomous items provide no information concerning which score levels are responsible for the DIF effect, the investigation of DIF within the context of each score level could provide a more comprehensive understanding of DIF in polytomous items. The investigation of the DIF effect associated with each score level can be accomplished using the framework of differential step functioning (DSF; Penfield, 2006, in press). The framework of DSF is founded on the concept of the step function. For a polytomous item having  $r$  ordinal score levels, the probability of observing each score level, given a particular level of ability, is determined by a set of  $J = r - 1$  step functions. Each of the  $J$  step functions specifies the probability of successfully advancing (i.e., stepping) from a lower score level to a higher score level as a function of ability. For example, consider a four-level item with score levels 0, 1, 2, 3. The three step functions underlying this item can be defined as follows: (a) the first step function specifies the probability of advancing from a 0 to a 1, 2, or 3 as a function of ability; (b) the second step function specifies the probability of advancing from a 0 or 1 to a 2 or a 3 as a function of ability; and (c) the third step function specifies the probability of advancing from a 0, 1, or 2 to a 3 as a function of ability. This example defines the step functions using a cumulative approach, but other approaches are also possible (Agresti, 1990).

The probability associated with each score level of a polytomous item is determined by the parameters of the  $J$  step functions underlying the polytomous response process. Therefore, the study of between-group differences in the probability of each score level of a polytomous item can be reduced to a study of the between-group difference in each of the parameters defining each of the  $J$  step functions. If no between-

group differences in the parameters of the  $J$  step functions exist, then it must be the case that no DIF exists. However, if a between-group difference does exist in the parameters defining one or more of the  $J$  step functions, then DIF will exist in the polytomous item. The framework of DSF is based on measuring between-group differences in the parameters of the step functions underlying the polytomous response variable.

The framework of DSF provides the DIF analyst with several potential advantages over omnibus measures of DIF. First, tests of DSF can be more powerful than net tests of DIF when the magnitude and/or sign of the DSF effect vary across the steps underlying the polytomous response variable (Penfield, 2006, in press). In the extreme case where the sign of the DSF effect changes across the steps (i.e., is positive for one step, but negative for another), the power of tests of DSF have been shown to be nearly twenty times that of omnibus net DIF statistics (i.e., a power of .045 for the net DIF statistic compared to a power of .85 for the test of DSF; e.g., reported in Penfield, 2006). A second advantage of the DSF framework is that it allows the DIF analyst to pinpoint precisely which score levels (or steps) are responsible for an observed DIF effect. That is, if a polytomous item is flagged for DIF, then the analysis of DSF can be used to isolate the score levels that require further review with respect to content and potential scoring problems, and ultimately inform the factors involved in causing the DIF. In light of the increasing interest in understanding the causes of DIF (Bolt, 2000; Douglas, Roussos, & Stout, 1996; Gierl & Khaliq, 2001; Oshima, Raju, Flowers, & Slinde, 1998; Scheuneman, 1987; Schmitt, Holland, & Dorans, 1993; Swanson, Clauser, Case, Nungester, & Featherman, 2002), the framework of DSF can serve as a useful tool in this endeavor.

Several statistical approaches have been proposed for the analysis of DSF: (a) an IRT approach (Cohen, Kim, & Baker, 1993) that is based on examining the step functions underlying the graded response model (GRM; Samejima, 1969, 1972); (b) a logistic regression approach (French & Miller, 1996); and (c) an odds ratio approach (Penfield, 2006, in press). While the IRT and logistic regression approaches have the appeal of a model-based assessment of DSF, their practical utility is hampered by: (a) the assumption of model fit that, if violated, can lead to an inflated Type I error rate (Bolt, 2002); (b) relatively large group sizes required for stable parameter estimation (French & Miller, 1996; Reise & Yu, 1990); and (c) knowledge of the models and appropriate software to run the models. In contrast, the odds ratio approach is not hampered by requirements of model fit and large sample sizes, and can be run using the easily accessible (point-and-click) DIFAS<sup>1</sup> computer program (Penfield, 2005).

While the concept of DSF has been introduced in the measurement literature as a mechanism to examine between-group measurement equivalence in polytomous items, these published accounts have provided limited information concerning precisely how the results of a DSF analysis can be used in conjunction with a DIF analysis in assessing the measurement equivalence of polytomous items and determining the causes of any identified lack of equivalence. In this paper we build on the previously published research on DSF by providing a comprehensive account of how the consideration of DSF can

---

<sup>1</sup> DIFAS is a windows-based point-and-click program that can conduct a variety of contingency table analyses pertaining to DIF in dichotomous and polytomous items, differential test functioning (DTF) for tests containing dichotomous and/or polytomous items, and DSF. The program reads in data from a space-, comma-, or tab-delimited file, and produces output in a text box contained in the graphical user interface that can be saved or cut-and-pasted into other word processing documents. DIFAS can accommodate up to 1000 items, and the number of cases is limited only by the resources of the computer running the analysis. DIFAS, and its corresponding manual, can be obtained free of charge by sending an e-mail to Randall Penfield.

enhance the analysis of measurement equivalence over that provided solely by traditional omnibus measures of DIF in polytomous items. Specifically, in this paper we aim to: (a) provide a structured taxonomy of the different forms of DSF that can occur, (b) describe how knowledge of the form of DSF can be strategically used in guiding item content review and/or revision, (c) present objective criteria for determining the form of the DSF, and (d) present guidelines for using the results of DIF and DSF in tandem to best assess lack of measurement equivalence in the items of a test for two or more groups. To demonstrate how these concepts can be applied to real testing situations, we provide an illustrative example of a DSF analysis on data obtained from a standardized science test administered in a large urban school district.

### **A Taxonomy of DSF**

DSF can take on several forms depending on the pattern of DSF effects across the  $J$  steps of the polytomous item under investigation. The specific pattern of DSF effects can play an important role in guiding the DIF analyst in interpreting the results of a DIF analysis, identifying the possible causes of the DIF effect, and ultimately making decisions about item revision and removal. In this section we propose a taxonomy of DSF that is intended to assist test developers identify and interpret the causes of DIF in polytomous items. The taxonomy categorizes the DSF according to two dimensions: (a) the locus of the potential causes of the DSF effect (i.e., a causal influence that manifests itself at the item level in contrast to a causal influence that targets particular steps of the polytomous response variable); and (b) the consistency of the causal influence across the steps manifesting a DSF effect. A summary of the taxonomy is presented in Table 1.

The first dimension of the DSF taxonomy distinguishes between *pervasive* and *non-pervasive* DSF. Pervasive DSF corresponds to the situation where all  $J$  steps display a substantial DSF effect, and thus the DSF effect is pervasive across all response options. The presence of pervasive DSF can suggest that the factor causing the DIF is exerting its influence at the level of the item, rather than at the level of a particular step. For example, consider the polytomous ratings of a writing task pertaining to a particular topic. The presence of pervasive DSF in the obtained scores to this item may indicate that the DIF effect is caused by a property of the general topic of the prompt as opposed to more specific aspects of the task that are unique to individual steps underlying the rating process (e.g., successfully integrating particular information).

Non-pervasive DSF corresponds to the situation whereby some, but not all, steps display a substantial DSF effect. The presence of non-pervasive DSF suggests that the factor causing the DIF may be localized to just one step, or possibly a few steps. For example, the presence of non-pervasive DSF observed in a writing task may indicate that the DIF is attributable to properties of one or more of the individual components of the task, rather than a property of the prompt itself. Examining the categories that define the steps displaying substantial DSF can guide the analyst in determining the location of the causes of DIF. For example, non-pervasive DSF observed in just the first step suggests that the cause of the DIF likely resides in second lowest score level because one group of examinees is experiencing a relative difficulty in making the transition from the lowest score level to a higher score level (assuming a cumulative step function is used).

The second dimension of the DSF taxonomy concerns the consistency of the DSF effect across the impacted steps, and distinguishes among *constant*, *convergent*, and

*divergent* DSF forms. Constant DSF concerns the situation whereby the steps displaying a substantial DSF effect are relatively equal in magnitude and sign. The situation of constant pervasive DSF provides strong evidence that the DIF effect is attributable solely to an item-level property. Constant non-pervasive DSF suggests that the cause of the DIF effect is not necessarily an item-level property, but is a property that is contained in all steps displaying a substantial DSF effect. In this situation, the DIF analyst should focus attention on properties that are shared across the score levels defining the impacted steps.

Convergent DSF represents the situation whereby all steps displaying a substantial DSF effect have the same sign, but not the same magnitude. The presence of convergent DSF provides evidence that while the properties causing the DIF are favoring the same group across all impacted steps, the causal properties are manifested differentially across the impacted steps. In this situation, it is possible that a single cause of the DIF resides at the item level and that the impact of the cause varies depending on the properties of each score level. Or, it is possible that there is more than one causal property at work across different score levels. As a result, the interpretation of the precise causes associated with convergent DSF can be challenging, particularly if numerous steps are impacted (i.e., exhibit substantial DSF). The DIF analyst is responsible for examining the properties of the score levels defining the steps in an effort to identify whether there is a single cause with a differential impact across the impacted steps or if there are multiple causes impacting separate steps.

In contrast to convergent DSF, *divergent* DSF represents the situation of different steps displaying DSF effects that have different signs. The presence of *divergent* DSF indicates that the relative advantage shifts between groups across the steps; one group

experiences a relative advantage on one or some steps, and the other group experiences a relative advantage on one or more other steps. The presence of divergent DSF provides strong evidence that the causes of the DSF are factors that are specific to individual steps, and thus subsequent content analysis used to identify the causes of the DSF effects should be targeted to the score levels defining the relevant steps. Furthermore, divergent DSF suggests that there exists more than one causal property of DIF, and that the different causal properties lead to a relative advantage of different groups. It is relevant to note that the presence of divergent DSF poses a particular problem for net DIF statistics because the positive and negative DSF effects can cancel to result in a zero (or near zero) net effect. As a result, net DIF tests are relatively insensitive to the condition of divergent DSF. Global DIF tests, however, consider the unsigned aggregation of DSF effect, and thus are relatively sensitive to divergent DSF.

### **Objective Criteria for Determining the DSF Form**

Identifying the form of DSF (i.e., pervasive, non-pervasive, constant, convergent, and divergent) is dependent on establishing the sign and magnitude of DSF effect at each step. Determining the sign of the DSF effect is relatively straight forward. However, determining whether the pattern of DSF effect magnitudes are relatively constant, convergent, or divergent, is less objective absent of a set of criteria to guide the analyst in determining the pattern of DSF effect magnitudes. Some analysts with sufficient experience examining the results of DIF and DSF analyses will, no doubt, develop an intuitive sense as to the form of the DSF pattern based on the DSF effect estimates. Other analysts, however, will benefit from an objective set of rules to guide them in determining the specific DSF form.

To develop a set of rules to help define the DSF form, it is first necessary to develop a scheme for categorizing the DSF effect with respect to its sign and magnitude. A set of rules can be developed based on any of the methodologies used to estimate the DSF effect, including IRT, logistic regression, and odds ratio approaches. In this paper we present a DSF effect classification scheme that is based on the common log-odds ratio approach (Penfield, 2006, in press) because of the widespread use of the odds ratio approach in conducting DIF analyses in dichotomous and polytomous item formats. The common log-odds ratio associated with the  $j$ th step is computed using

$$\hat{\lambda}_j = \ln \left[ \frac{\sum_{k=1}^m \frac{A_{jk} D_{jk}}{T_k}}{\sum_{k=1}^m \frac{B_{jk} C_{jk}}{T_k}} \right], \quad (1)$$

where  $A_{jk}$  represents the number of reference group members at the  $k$ th stratum of ability (e.g.,  $k$ th total test score level) who successfully advanced at the  $j$ th step,  $B_{jk}$  represents the number of reference group members at the  $k$ th stratum of ability who did not successfully advance at the  $j$ th step,  $C_{jk}$  represents the number of focal group members at the  $k$ th stratum of ability who successfully advanced at the  $j$ th step,  $D_{jk}$  represents the number of focal group members at the  $k$ th stratum of ability who did not successfully advance at the  $j$ th step, and  $T_k$  represents the total number of reference and focal group members at the  $k$ th stratum of ability. Under the cumulative definition of the step function, advancing at the  $j$ th step corresponds to having an item response ( $Y$ ) that is equal to or greater than the  $j$ th score level. As a result,  $A_{jk}$  and  $B_{jk}$  are equal to the number of reference group members at the  $k$ th stratum for whom  $Y \geq j$  and  $Y < j$ , respectively. Similarly,  $C_{jk}$  and  $D_{jk}$  are equal to the number of focal group members at the  $k$ th stratum

for whom  $Y \geq j$  and  $Y < j$ , respectively. Additional information concerning the calculation of  $\hat{\lambda}_j$  and its estimated standard error is provided by Penfield (2006, in press).

The step-level common log-odds ratio estimator associated with the  $j$ th step ( $\hat{\lambda}_j$ ) shown in Equation 1 is identical in form to the Mantel-Haenszel common log-odds ratio ( $\hat{\lambda}_{MH}$ ; Mantel & Haenszel, 1959) widely used in DIF detection for dichotomous items (Camilli & Shepard, 1994; Holland & Thayer, 1988). As a result, it is justifiable to interpret the magnitude of the step-level log-odds ratio using a metric that is equivalent to that used for interpreting  $\hat{\lambda}_{MH}$  in the study of DIF in dichotomous items. A widely used scheme for interpreting the magnitude of  $\hat{\lambda}_{MH}$  is the ETS classification scheme (Zieky, 1993). Under the ETS classification scheme,  $|\hat{\lambda}_{MH}| < 0.43$  corresponds to a small DIF effect,  $0.43 \leq |\hat{\lambda}_{MH}| < 0.64$  corresponds to a medium DIF effect, and  $|\hat{\lambda}_{MH}| \geq 0.64$  corresponds to a large DIF effect (the full ETS scheme also includes significance tests to classify the DIF severity, but we focus here only on the size of the effect for the purposes of the DSF analysis).

The ETS classification scheme described above can be applied to the step-level common log-odds ratio estimators used in a DSF analysis. Specifically, the DSF effect can be categorized as small, medium, or large by applying the ETS classification scheme criteria to the step-level common log-odds ratio for the  $j$ th step ( $\hat{\lambda}_j$ ). In applying the ETS scheme, however, it is necessary to consider the sign of the DSF effect in order to distinguish between convergent and divergent forms of DSF. As a result, we propose using the following classification scheme for the DSF effect of the  $j$ th step

Small DSF Effect (S)	$ \hat{\lambda}_j  < 0.43$
Medium Negative DSF Effect (M-)	$-0.64 < \hat{\lambda}_j \leq -0.43$
Medium Positive DSF Effect (M+)	$0.43 \leq \hat{\lambda}_j < 0.64$
Large Negative DSF Effect (L-)	$\hat{\lambda}_j \leq -0.64$
Large Positive DSF Effect (L+)	$\hat{\lambda}_j \geq 0.64$

We want to stress that this DSF classification scheme is consistent with the ETS scheme used with dichotomously scored items; the A, B, and C categories of the ETS scheme correspond to small, medium, and large categories in the DSF classification scheme. As such, the classification of DSF effect can be conducted in a fashion that is similar to that done with dichotomously scored items.

It is relevant to note that the DSF classification scheme described above differentiates between positive and negative DSF effects for the medium and large categories, but not for the small category. The rationale for not distinguishing between positive and negative small effects is based on the conceptualization of a small DSF effect as being representative of a negligible DSF effect. That is, a small DSF effect is small enough that, for the purposes of defining the DSF form, it is considered the absence of a DSF effect. As a result, the sign of the small DSF effect is irrelevant to the form of the DSF. Rather, the presence of a small DSF effect for one or more steps plays a defining role in distinguishing between pervasive and non-pervasive DSF forms.

Based on the five-category DSF classification scheme proposed above, it is possible to systematically define each possible form of DSF described in the DSF taxonomy presented in Table 1. For example: (a) if all steps are categorized as having a

medium positive DSF effect, then the DSF is pervasive and constant in form; (b) if one step is categorized as having a medium negative DSF effect but all other steps are categorized as having a large positive DSF effect, then the DSF is pervasive and divergent in form; and (c) if all steps are categorized as having a small DSF effect except for two steps, both categorized as having a large negative DSF effect, then the DSF is non-pervasive constant in form. Table 2 presents a description of each of the possible DSF forms defined in the taxonomy with respect to the five DSF effect categories described above.

The DSF effect categorization scheme presented above is intended to be a useful mechanism for classifying the form of DSF. This categorization scheme, however, can be adjusted depending on the needs of the test developer. For example, a more liberal definition of a small DSF effect may be treated as one for which  $|\hat{\lambda}_j| < 0.20$ , and the corresponding definitions of medium and large DSF effects could be modified accordingly. It is also possible to incorporate hypothesis tests in the classification scheme such that the classification of medium and large DSF effects is contingent not only on the size of the DSF effect, but also on the rejection of one or more hypotheses concerning the DSF effect (i.e., the null hypothesis that the DSF effect equals zero). It has been our experience, however, that with moderate group sizes (i.e.,  $> 300$  per group) the categorization each DSF effect is typically determined by the magnitude DSF effect rather than the outcome of a hypothesis test (due to the relatively high power involved in the hypothesis tests with moderate or large group sizes), and thus categorizing the level of DSF according to the DSF effect size alone is likely sufficient, given adequate group sizes.

## **Guidelines for Using DSF and DIF in Combination**

The assessment of DIF for polytomous items typically yields a single item-level index, or statistical test, of measurement equivalence. If lack of measurement equivalence exists in a polytomous item, however, DIF statistics provide no information concerning where among the score levels the effect exists. This is where the utility of DSF framework is realized; DSF can shed light on which score levels are displaying the measurement nonequivalence. However, if the analysis of DSF is to be used, it is important to establish how the results of the DSF analysis should be used in relation to the results of the DIF analysis. Should DSF be used only after observing a statistically significant item-level DIF effect? Should the framework of DSF replace the DIF paradigm for polytomous items, such that all DIF analyses for polytomous items are reduced to step-level analyses? Or, should some other combination of DIF and DSF be the standard choice? This section serves to address these questions and ultimately provide a set of guidelines for using the DIF and DSF frameworks in a complementary manner.

In examining how the frameworks of DIF and DSF can be most effectively utilized, it is important to first establish the relative strengths and weaknesses of each framework in assessing measurement equivalence in polytomous items. The DIF framework has the weakness of providing no information concerning which score levels are involved in the DIF effect. In contrast, the DSF framework provides comprehensive information about which score levels are responsible for an item-level DIF effect. Despite the limited information provided by the DIF framework, tests of DIF have the potential to be more powerful than tests of DSF because DIF analyses collapse across data contained at all  $J$  steps. To summarize, the DIF framework is expected to be more sensitive in

detecting measurement non-equivalence, but the DSF framework is expected to be more informative in determining what the measurement non-equivalence actually looks like.

In light of the strengths and limitations of the DSF and DIF frameworks, we recommend beginning any study of measurement equivalence in polytomous items with a statistical test of the null hypothesis of no DIF using a combination of global and net DIF tests. Previous research by Penfield (2006, in press) and Wang & Su (2004) showed that: (a) global DIF tests, such as the GMH test (Somes, 1986; Zwick et al., 1993) and the SSL<sup>2</sup> test (Penfield, in press), are more powerful than net DIF tests when the DSF effect was not constant across all steps (i.e., any DSF form other than pervasive constant); and (b) net DIF tests, such as Mantel's chi-square (Mantel, 1963; Zwick et al., 1993) and the cumulative common log-odds ratio test (Penfield & Algina, 2003), are more powerful than global DIF tests when the DSF effects are constant across all steps (i.e., DSF form is pervasive constant). As a result, a comprehensive DIF analysis for polytomous items should include at least one global DIF test (i.e., GMH test or SSL test) and one net DIF test (i.e., Mantel's chi-square or the cumulative common log-odds ratio test). Naturally, there are numerous other global and net DIF tests (discussed in the Introduction) that can be substituted for those mentioned here. If the null hypothesis of no DIF is accepted for both the global and net DIF tests, then it can be concluded that measurement equivalence exists and the analysis can be terminated.

---

<sup>2</sup> The SSL test of global DIF is based on testing the null hypothesis of no DSF at each of the  $J$  steps using a Bonferroni-adjusted Type I error rate, and rejecting the null hypothesis of no DIF if the null hypothesis of no DSF is rejected for one or more of the  $J$  steps. Penfield (in press) proposed testing the null hypothesis of no DSF at each step using the test statistic  $z = \hat{\lambda}_j / SE(\hat{\lambda}_j)$ , which is asymptotically distributed as standard normal (Hauck, 1979; Penfield, in press).

If the null hypothesis of no DIF is rejected using either of the global or net tests of DIF, then a thorough DSF analysis should be conducted. At a minimum, the DSF effect at each of the  $J$  steps should be estimated and interpreted with established criteria, such as those described in the previous section. Based on the sign and magnitude of the DSF effects across the  $J$  steps, the form of the DSF can be established using the taxonomy displayed in Table 1, and the resulting DSF form can provide guidance to the analyst concerning precisely where investigations of item content should be targeted.

### **An Illustrative Example**

To illustrate the use of DSF and the proposed DSF taxonomy, we now present an application of these methods to a science test administered to 4<sup>th</sup> grade students in a large urban school district. The test was developed as part of a study examining the impact of professional development on the science achievement of English language learners in the context of the impending high-stakes testing and accountability policy in science under the No Child Left Behind (NCLB) Act. The test measured key science concepts related to the 4<sup>th</sup> grade curriculum units including measurement, matter, water cycle, weather, energy, force and motion, and processes of life. The items of the test were a combination of project-developed items and publicly released items from the National Assessment of Educational Progress (NAEP), Third International Mathematics and Science Study (TIMSS), and Florida Comprehensive Assessment Test (FCAT). Item formats included multiple choice, short answer, and extended response. The test consisted of 25 items, of which four were scored polytomously. Possible test scores ranged from 0 to 31. Because the test was administered in schools composed primarily of Hispanic and Black students

(including many Haitian and Caribbean immigrants), the measurement equivalence of the items of the test across the two groups was of primary interest.

The analyses reported here were based on a comparison of 291 Hispanic (coded as the reference group) and 357 Black (coded as the focal group) students. The mean and standard deviation of the total test score were similar for both groups (for the Hispanic group  $M = 14.95$  and  $SD = 4.62$ , and for the Black group  $M = 13.24$  and  $SD = 4.10$ ). The estimated reliability of the observed scores (using Cronbach's alpha) was 0.74, which admittedly is not very high for high stakes tests of achievement but was deemed to be high enough for the didactic purposes of this exposition. The following analyses were conducted: (a) a DSF analysis was conducted for the polytomously scored items using the step-level common log-odds ratio ( $\hat{\lambda}_j$ ) as computed using the DIFAS computer program (Penfield, 2005); (b) the DSF form was identified using the taxonomy and categorization schemes presented in Tables 1 and 2; (c) a test of the null hypothesis of no net DIF was conducted using Mantel's chi-square test with a Type I error rate of .05 (using DIFAS); and (d) a test of global DIF for the polytomous items was conducted using the SSL test using a familywise Type I error rate of .05 such that the step-level tests were conducted using  $z = \hat{\lambda}_j / SE(\hat{\lambda}_j)$  evaluated with a Bonferroni-adjusted Type I error rate of  $.05/J$  (i.e., .0167 for  $J = 3$  and .025 for  $J = 2$ ), as described by Penfield (in press). As advocated by Zwick et al. (1993) the stratifying variable used in all DSF and DIF analyses was the summated test score across all dichotomous and polytomous items of the test.

Table 3 presents the DIF and DSF results for the polytomous items of the test. Two items (Items 1 and 3) displayed a nonsignificant test of the global DIF and small DSF effects. The remaining two items (Items 2 and 4) displayed DSF effects that were

either medium or large in magnitude. Item 2 did not have a significant net or global DIF test, but did display a medium negative DSF effect for the third step ( $\hat{\lambda}_3 = -0.45$ ). While this DSF effect was not statistically different from zero, its medium magnitude indicates a potential biasing property of the uppermost response categories. Item 4 demonstrated a significant global test of DIF, a medium positive DSF effect for the first step ( $\hat{\lambda}_1 = 0.61$ ), a large negative DSF effect for the second step ( $\hat{\lambda}_2 = -1.63$ ), and a small negative DSF effect for the third step ( $\hat{\lambda}_3 = -0.22$ ). This pattern of DSF effects characterizes a non-pervasive divergent DSF form. The results suggest that transition from the lowest score level to the higher response categories was relatively more difficult for Black respondents than Hispanic respondents (i.e., the second score level may contain a biasing factor against the Black group). In contrast, the transition from either of the lowest two response categories to the uppermost response categories was more difficult for Hispanic versus Black respondents (i.e., the third and fourth response categories may contain a biasing factor against the Hispanic group).

To further illustrate how DSF can be used to identify causes of DIF, let us explore the potential causes of the DSF effects observed in Item 4. This item consisted of a list of eight living things (i.e., gorilla, parrot, snake, earthworm, jellyfish, sponge, fish, fly) and asked the examinee to divide them into two groups according to an important physical characteristic, to describe the primary physical characteristic used in the classification, and to describe a potential second characteristic that could have been used to classify the living things. The highest score level corresponded to an acceptable categorization and a clear description of the primary and second physical characteristic used in the classification. The second highest score corresponded to an acceptable categorization and

primary physical characteristic for the classification without providing a valid second characteristic for classification. The next highest score corresponded to an acceptable categorization without citing the correct primary or second physical trait used in the categorization. The lowest score corresponded to the absence of any acceptable categorization.

The positive DSF effect in the first step of Item 4 ( $\hat{\lambda}_1 = 0.61$ ) indicates that the transition from the lowest score level (providing no acceptable categorization) to a higher score level (i.e., providing an acceptable categorization but not necessarily providing a valid primary or second physical trait) was relatively easier for the Hispanic group than the Black group. The precise cause of this is unknown, but it may be that the living things presented in the item had greater familiarity to the Hispanic group than the Black group. The negative DSF effect in the second step of Item 4 ( $\hat{\lambda}_2 = -1.63$ ) indicates that the transition from the second lowest score level (i.e., being able to provide an adequate classification without providing a valid physical trait) to a higher score level (i.e., being able to provide at least a valid primary trait) was relatively easier for the Black group. This may indicate that communicating the physical trait used in classifying living things poses a particular problem for individuals for whom English is not the first language. That is, the language demands required to explain the trait used in creating the classification may be large, thus generating a biasing factor in the item.

The results of the DSF analyses presented in this section demonstrated examples of non-pervasive DSF, but did not yield any examples of pervasive DSF. These results should not, however, be taken as evidence that pervasive DSF will not exist in practice. Applications of DSF methodology to real testing data have not been presented previously

in the literature, and as a result there is little knowledge of the prevalence of different forms of DSF in real data.

### **Discussion**

Differential Step Functioning (DSF) is a framework for examining measurement equivalence within each step of a polytomous response variable. Using DSF in coordination with omnibus DIF statistics can yield a more comprehensive assessment of measurement equivalence than DIF analyses alone, particularly when the magnitude and/or sign of the DSF effect changes across steps. Additionally, examination of the location and form of the DSF effects allows the analyst to pinpoint the locus of the potential biasing factor (i.e., an item-level factor versus a factor that is restricted to particular score levels) and, where applicable, determine which score levels are implicated in the observed DIF effect. For these reasons, DSF can provide valuable aid to researchers hoping to understand the causes of DIF in polytomous items.

In using the DSF framework to identify the causes of the DIF, it is important to note that DSF may not be well suited for every form of polytomous item. Testlets, for example, are not appropriate for the DSF approach because the results of the DSF analysis cannot inform where (i.e., which multiple-choice item of the testlet) the problem is arising. That is, DSF will identify whether measurement equivalence exists with respect to each number of correct responses in the testlet, but not within particular testlet items. In this situation, the DIF analyst is better off running individual dichotomous DIF analyses on each of the multiple-choice items contained in the testlet. Similarly, polytomous items for which the score corresponds to a frequency of successfully

completed tasks (i.e., the number of correctly labeled aspects of a graph) are not well suited for the DSF framework for reasons similar to that of the testlet.

The study of DSF, and its application to the analysis of the causes of DIF, is in its infancy. As a result, numerous questions concerning the properties of DSF measures remain unanswered in the measurement literature. One such question is how the methods available for studying DSF (i.e., IRT, logistic regression, and odds ratio approaches) compare with respect to their statistical properties, both in the context of empirical simulation studies and in their application to real data sets. A related question is how alternative methods for assessing DSF might improve upon currently available methodology. For example, the SIBTEST procedure (Shealy & Stout, 1993), which can reduce the inflation in the Type I error rate caused by a substantial between-group difference in ability distribution, may be applied to the analysis of DSF and provide more robust tests of DSF under non-ideal conditions. In addition, future research may benefit from better understanding the link between the results of a DSF analysis and the causes of DIF, particularly in the context of the DSF taxonomy proposed in this paper. Relevant issues include: (a) understanding the forms of DSF that most frequently occur in actual polytomous items; (b) understanding which forms of DSF are associated with particular types of polytomous items (i.e., writing tasks, performance tasks); (c) improving the interpretation of the location of the factors causing the DIF effect based on the results of a DSF analysis; and (d) developing graphical displays of DSF effects that can aid the test developer in diagnosing potential problems with item content and/or rater scoring.

## References

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: APA.
- Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.
- Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.
- Bolt, D. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.
- Cohen, A. S., Kim, S. -H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350.

- Dorans, N.J., & Schmitt, A.P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive assessment: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum.
- Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465-484.
- Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.
- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164-187.
- Hauck, W. W. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics, 35*, 817-819.
- 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: Erlbaum.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ:

- Erlbaum.
- Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.
- Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.
- Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Psychological Measurement, 11*, 353-369.
- Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150-151.
- Penfield, R. D. (2006, April). *A nonparametric method for assessing differential step functioning in polytomous items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353-370.

- Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C.R. Rao (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 125-167). New York: Elsevier.
- Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15.
- Penfield, R. D. (in press). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*.
- Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.
- Roussos, L.A., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-115). Thousand Oaks: Sage.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (No. 17). Psychometric Monograph.
- Samejima, F. (1972). *A general model for free response data* (No. 18). Psychometric Monograph.
- 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., 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: Lawrence Erlbaum.
- Shealy, R. T., & Stout, W. F. (1993). 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*, 197-239.
- Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, *40*, 106-108.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53-75.
- Wang, W., -C., & Su, Y., -H. (2004). Factors influencing the Mantel and Generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*, 450-480.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.

TABLE 1  
Taxonomy of DSF Form

Consistency (Consistency of Cause of DIF)	Pervasiveness (Locus of Cause of DIF)	
	Pervasive	Non-Pervasive
Constant	All steps display a DSF effect equal in magnitude and sign	Only one or a few steps display a DSF effect equal in magnitude and sign across affected steps
Convergent	All steps display a DSF effects having the same sign, but different magnitudes	Only a few steps display a DSF effects having the same sign, but different magnitudes
Divergent	All steps display a DSF effect, but the signs vary across steps	Only a few steps display a DSF effect and the signs vary across steps

TABLE 2  
Description of the DSF Form Based on a DSF Effect Categorization Scheme

DSF Form	Description	Example for a 3-Step Item		
		Step 1	Step 2	Step 3
No DSF	All steps categorized as S	S	S	S
Pervasive Constant	All steps categorized entirely as M-, or entirely as M+, or entirely as L-, or entirely as L+	M+	M+	M+
Pervasive Convergent	All steps categorized entirely as M- and L- OR All steps categorized entirely as M+ and L+	M-	M-	L-
Pervasive Divergent	No steps categorized as S AND At least one step categorized as M- or L- AND At least one step categorized as M+ or L+	M+	L+	L-
Non-Pervasive Constant	At least one step categorized as S AND All other steps categorized entirely as M-, or entirely as M+, or entirely as L-, or entirely as L+	S	L-	L-
Non-Pervasive Convergent	At least one step is categorized as S AND All other steps categorized as a mixture of M- and L- or a mixture of M+ and L+	S	M+	L+
Non-Pervasive Divergent	At least one step is categorized as S AND At least one step is categorized as M- or L- AND At least one step is categorized as M+ or L+	S	L+	L-

*Note.* In the descriptions and examples provided, S = small DSF effect, M- = medium negative DSF effect, M+ = medium positive DSF effect, L- = large negative DSF effect, and L+ = large positive DSF effect.

TABLE 3  
Results of the DIF and DSF Analyses for the Fourth Grade Test

Polytomous Item	DSF Measures				DIF Measures	
	Step 1 ( $\hat{\lambda}_1$ )	Step 2 ( $\hat{\lambda}_2$ )	Step 3 ( $\hat{\lambda}_3$ )	DSF Form	Global DIF (SSL Test)	Net DIF (Mantel)
1	0.03 (0.22) S	0.32 (0.23) S	NA	No DSF	Accept	Accept $\chi^2 = 0.98$
2	0.02 (0.19) S	-0.33 (0.34) S	-0.45 (0.55) M-	Potential Non-Pervasive	Accept	Accept $\chi^2 = 0.29$
3	0.08 (0.21) S	-0.30 (0.31) S	NA	No DSF	Accept	Accept $\chi^2 = 0.03$
4	0.61* (0.19) M+	-1.63* (0.52) L-	-0.22 (1.35) S	Non-Pervasive Divergent	Reject	Accept $\chi^2 = 1.95$

*Note.* Standard errors are reported in the brackets. Asterisks indicate statistical significance at the appropriate Bonferroni-adjusted Type I error rate for the DSF effects (i.e., .0167 for items containing 3 steps).