Journal of Applied Psychology
2010, Vol. 95, No. 4, 728–743

0021-9010/10/$12.00 DOI: 10.1037/a0018966

A Taxonomy of Effect Size Measures for the Differential Functioning of Items and Scales

Adam W. Meade
North Carolina State University

Much progress has been made in the past 2 decades with respect to methods of identifying measurement invariance or a lack thereof. Until now, the focus of these efforts has been to establish criteria for statistical significance in items and scales that function differently across samples. The power associated with tests of differential functioning, as with all significance tests, is affected by sample size and other considerations. Additionally, statistical significance need not imply practical importance. There is a strong need as such for meaningful effect size indicators to describe the extent to which items and scales function differently. Recently developed effect size measures show promise for providing a metric to describe the amount of differential functioning present between groups. Expanding upon recent developments, this article presents a taxonomy of potential differential functioning effect sizes; several new indices of item and scale differential functioning effect size are proposed and illustrated with 2 data samples. Software created for computing these indices and graphing item- and scale-level differential functioning is described.

Keywords: measurement invariance, differential functioning, DIF, invariance testing

During the past three decades, methods of detecting a lack of invariance across samples have advanced rapidly. This is true of advances both in the use of confirmatory factor analytic methods (Vandenberg & Lance, 2000) and in item response theory methods (e.g., Raju et al., 2009). The efficacy of several such procedures to determine statistical significance of a lack of invariance has been supported via simulation work (e.g., Meade & Lautenschlager, 2004; Stark, Chernyshenko, & Drasgow, 2006). These methodological advances have focused almost exclusively on the determination of statistical significance of the differences in the psychometric properties of a given measure across samples. On the other hand, only recently has significant progress been made in establishing an index of effect size of a lack of invariance.

In this study, an item response theory (IRT) framework is used to establish a series of effect size indices of differential functioning (DF; i.e., a lack of invariance) at the item and scale levels. A taxonomy featuring key dimensions of potential DF effect sizes is proposed, and existing DF effect size measures are discussed within the context of this framework. Accessibility of methods of computing DF effect size measures is an important determinant of widespread utilization of such methods. As such, a computer program developed to produce DF effect size indices and graphs is described and illustrated using two data samples.

Invariance and DF

Measurement invariance refers to the extent to which the psychometric properties of a measure (e.g., scale, survey, test) are constant across samples (e.g., groups of respondents, time periods). Recently, tests of invariance, or DF, have figured more prominently in evaluations of the suitability of measures, as invariance must be present before meaningful comparisons in observed data can be made (Horn & McArdle, 1992; Raju, Laffitte, & Byrne, 2002; Vandenberg, 2002). Invariance may be particularly important in organizational contexts such as employee selection, in which it is imperative that a measure or test not function differently for persons belonging to different gender, racial, or other demographic groups (Society for Industrial and Organizational Psychology, 2003). Invariance is also important in other aspects of organizational research, such as comparing survey results across cultures or countries (e.g., Ryan, Horvath, Ployhart, Schmitt, & Slade, 2000), assessing change in satisfaction (e.g., Wu, Chen, & Tsai, 2009) and other employee attitudes over time, comparing performance ratings across different rater groups (Facteau & Craig, 2001), and comparing tests administered via paper and pencil and via the Internet (Meade, Michels, & Lautenschlager, 2007).

Many methods have been proposed to identify significant DF. Although a full review of such methods is well beyond the scope of this study, two methods have emerged as particularly promising for use in organizational research: the *likelihood ratio test* (LRT; Thissen, Steinberg, & Wainer, 1988, 1993) and the *differential functioning of items and tests* (DFIT) framework (Flowers, Oshima, & Raju, 1999; Raju, van der Linden, & Fleer, 1995). Both the LRT and DFIT have been shown to be effective at detecting DF (Bolt, 2002; Flowers et al., 1999; Meade & Lautenschlager, 2004; Meade, Lautenschlager, & Johnson, 2007; Raju et al., 2009; Stark et al., 2006). However, one considerable disadvantage is that the LRT has high power to detect even very small differences in item functioning when sample sizes are large (see Appendix A for a small simulated example). This is unfortunate, as large samples are generally recommended when estimating IRT models (Reise & Yu, 1990). The sampling distributions of DFIT are also affected by sample size (Meade, Lautenschlager, & Johnson, 2007).

Special thanks to S. Bart Craig, Scott Tonidandel, and Alan Mead for their feedback and guidance on an earlier draft of this article.

Correspondence concerning this article should be addressed to Adam W. Meade, Department of Psychology, Campus Box 7650, North Carolina State University, Raleigh, NC 27695-7650. E-mail: awmeade@ncsu.edu

The primary goal in examining invariance is to determine whether or not the items and scales function significantly differently for different samples (e.g., racial groups) or whether the differences in item parameters can be attributed to sampling error. This focus is understandable and parallels that of statistics developed for the comparison of group means, in that *t* tests and analysis of variance were necessary developments prior to the establishment of effect size measures of group differences such as Cohen's *d* (Cohen, 1988). However, a question of interest often is not just whether significant DF exists but how much DF exists.

Imagine a scenario in which a researcher has access to a large sample of responses (e.g., $N = 2{,}000$) to a personality test used for selection. After careful analyses, the researcher reports and submits for publication the results of a *t* test finding that men and women statistically differ in their responses to the personality measure. The researcher provides no further comment regarding which group had higher test scores or how large the difference was. In this hypothetical example, reviewers and readers with even a rudimentary understanding of statistics would want to know the mean and standard deviation of the test scores in each group, given the very high power of the *t* test. Though this example may seem absurd, it is nearly identical to the current situation witnessed in invariance research. Most typically invariance is found or not found, and this result is dutifully reported. There is seldom any indication of which group is "favored" by the differences in measurement properties or how much of an effect the DF exerts on test scores for each group. Rarely are implications for organizational measurement made explicit in quantitative terms.

Significant DF indicates that a measure does not function equivalently in different samples. Some authors have indicated that extreme DF may represent a fundamental difference in the construct across samples (Horn & McArdle, 1992; Vandenberg & Lance, 2000).[1] However, it is also possible that the construct itself is equivalent across samples but items function in ways that are slightly different. In other words, DF is best thought of as a continuum rather than a dichotomous entity, just as are effect sizes of other statistics. At times, despite their statistical significance, minor differences in item functioning may have a very small impact on observed scores. Additionally, the power of DF analyses, like that of other statistics (e.g., correlations, mean differences), is dependent upon not only the effect size (amount of DF present) but also sample size and other considerations. Thus, as with other statistics, there are occasions for which statistical significance may be reached even though there is virtually no practical importance of the DF effect. Though tests of significance play an important role in determining whether differences in item properties are the result of sampling error, they are not sufficient for fully understanding the effect of DF on observed scores. Easily interpretable effect size indices for DF at both the item and scale levels are badly needed. Such indices allow researchers to decide whether they wish to alter the measure in some way, ignore practically inconsequential DF, or "correct" observed score differences using DF estimates.

Recently, Stark, Chernyshenko, and Drasgow (2004) discussed methods to translate DF into real-world outcomes, such as differences in selection ratios. Additionally, Stark et al. (2004) and Raju et al. (1995) provided initial development of some promising DF effect size indices. Below, several new DF effect size indices are developed and a taxonomy is proposed to classify the new and previous DF effect size indices.

Background

The DF effect size measures proposed here build upon a foundation of the concept of expected scores, which are easily computed given IRT item parameter and latent trait estimates. Expected scores can be computed on the basis of any IRT model that allows for computation of the probability of responding to any one response option. The current study focuses on polytomous data for two reasons: (a) most organizational surveys and noncognitive selection tests use a polytomous response format and (b) the indices and formulas described also directly apply to dichotomous data.[2]

For any level of theta, the expected score for item *i* can be computed as the sum of the probabilities of a response to each of the *m* response options times the value of that option (X_k):

$$ES_{s(\hat{\theta})i} = \sum_{k=1}^{m} P_{ik}(\hat{\theta})X_{ik} \tag{1}$$

where $ES_{s(\hat{\theta})i}$ is the expected score for respondent *s* (with a given estimated theta value) for item *i*, $P_{ik}(\hat{\theta})$ is the probability of responding to category *k* given the respondent's estimated theta level, and X_{ik} is the value of response option *k* (e.g., X_{im} would be 5 for an item with five response options). The expected score is similar to an item-level true score and has a potential range from the lowest response option (e.g., 1) to the highest (e.g., 5). Similarly, expected scale (or test) scores (ETS) for each respondent *s* is computed as the sum of the item expected scores (ESs) for person *s*:[3]

$$ETS_s = \sum_{i=1}^{j} ES_{si} \tag{2}$$

There is some consensus that DF is best evaluated with respect to the minority (focal) group (Flowers et al., 1999; Raju et al., 1995; Stark et al., 2004). In order to do so, one must estimate item parameters and link them to a common metric. This can be done via concurrent calibration, as is done using the LRT approach, or via separate calibration and linking using a routine such as Baker's (1995) modified version of the Stocking and Lord (1983) characteristic curve approach (cf. Raju et al., 2009). Only DF-free items should be used to link parameters if the characteristic curve approach is used (see Stark et al., 2004). Once linked to a common

[1] Most of these arguments have focused on the confirmatory factor analytic approach to examining invariance in cases in which the factor structure itself varies across groups (i.e., configural invariance).

[2] Dichotomous data should be scored 1 and 2 rather than the traditional 0 and 1 with the formulas described.

[3] Stark et al. (2004) used the term *test characteristic curve* (TCC) rather than *ETS.* Flowers et al. (1999) used the term *true score* (T). TCC, T, and ETS are different ways of referring to the same value. I prefer the term *ETS* over *T*, as it differentiates that the score is an expectation and also deviates from classical test theory uses of the term *true score.* I prefer ETS over TCC, as the term *TCCs* describes the graphical appearance of ETSs taken in aggregate. I believe ETS is clearer when discussing scores of individuals.

metric, each item is associated with two sets of item parameters, one that applies to the reference (majority) group, R, and one for the focal (minority) group, F. Latent trait theta estimates must also be computed for the focal group. Next, ESs can be computed for each member of the focal group by using focal group item parameters and also from reference group item parameters by using Equation 1. Plots of ESs when using these two different sets of parameters provide a visual indication of the extent to which DF is present (see Figure 1 for an example).

Taxometric Framework and DF Effect Size Measures

There are many different classes of potential effect size measures for DF. For example, there are R^2-based indices (Zumbo, 1999), model-impact-based indices (Kim, 2000; Wainer, 1993), indices based on estimates from observed scores (Dorans & Schmitt, 1991), and standardized differences in the metric of item parameters (Steinberg & Thissen, 2006). The taxonomy provided in this study focuses exclusively on effect size measures derived from ESs and ETSs for two reasons. First, the original metric of the observed scores is perhaps the most intuitive and easy to grasp for both experts and laypersons. Second, ESs and ETSs provide an intuitive summary of what may be complicated patterns of DF within and across items. Additionally, although other approaches (e.g., SIBTEST, Shealy & Stout, 1993; PolySIBTEST, Chang, Mazzeo, & Roussos, 1996) provide similar information, the discussion here is limited to parametric IRT-based models.

Several DF effect size indices follow directly from the item ESs and the scale ETSs. There are four primary dimensions upon which these DF effect size measures can be based. The first is whether the index is an item-level or scale-level indicator of effect size. Second, the index may use actual focal-group sample data or an assumed distribution to index the extent of DF. The third is the extent to which cancellation of DF is allowed to occur. As seen in the example ETS plot in Figure 1, DF need not consistently favor one group across all levels of the latent trait. Further, cancellation can take place in two ways: (a) within a given item, by averaging across focal group respondents, and (b) within persons across items for scale-level indices. At the item-level, DF can be nonuniform such that the focal group will have lower ESs than the reference group for some areas of theta but the opposite may be true for other regions of theta. Averaging across all respondents in the focal group allows these differences in expected scores to cancel within a single item, resulting in a small value for the DF

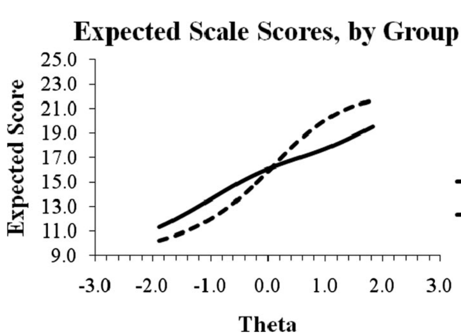

Figure 1. Example expected scale score plots for the focal group using both focal and reference group parameters. Ref = reference.

effect size index. Conversely, when the absolute value of these differences is taken, DF is not allowed to cancel at the item level. Similarly, at the scale level, DF may be allowed to cancel across items or not. The ETS is the sum of the expected item scores; thus, ETSs inherently allow cancellation of DF across items for each respondent. The mean difference in ETS across the sample goes further in allowing full cancellation across both items and person. On the other hand, taking the absolute value of the difference in ETS for each person prior to averaging allows for cancellation of DF across items but not across persons.

Last, the fourth dimension upon which effect sizes may vary in the taxometric framework is that they may be normed or left in the original metric. Expected item and scale scores are naturally in the same metric as observed scores. Differences in ESs can be left in the original observed score metric, or they can be normed to a standard deviation metric similar to Cohen's (1988) d (e.g., Stark et al.'s 2004 d_{DTF} index).

Development of multiple DF effect size indices and a taxonomy to classify these indices is important. First, each DF effect size provides slightly different, if generally overlapping, information about the magnitude and nature of the DF present in the item and test scores. The researcher can get a fuller understanding of the DF present by examining several indices rather than any one index alone. Additionally, by developing multiple indices simultaneously and providing a taxonomy for understanding how these indices relate to one another, researchers are presented with many options for summarizing the DF present in their measures.

Item-Level Indices

$ES_{(si|\hat{\theta},\gamma_F)}$ represents the ES for respondent s on item i, given that respondent's estimated theta level and the estimated focal group item parameters (γ_F). From this, the most basic index of DF for a sample of focal group respondents is the average, across the N respondents, of the difference in the respondents' ESs,

$$SIDS_i = \frac{\sum_{s=1}^{N}[ES_{(si|\hat{\theta},\gamma_F)} - ES_{(si|\hat{\theta},\gamma_R)}]}{N} \tag{3}$$

where SIDS_i represents the signed item difference in the sample for item i. Note that the SIDS_i index is conceptually similar to Raju et al.'s (1995) and Flowers et al.'s (1999) CDIF index. SIDS is also a polytomous extension of the dichotomous item SPD-θ index proposed by Camilli and Shepard (1994).

The SIDS index can be interpreted as the average difference in ESs across the sample of focal group respondents. An advantage of the SIDS index is that it is in the metric of ESs on the item, which is the same as that of the observed scores. Thus, a SIDS of −1.5 for an item with five response options would indicate that on average, focal group members in the current sample would be expected to score 1.5 points lower on the item than would reference group members with equal estimated trait scores. By expressing DF in the metric of observed scores, researchers can readily interpret the effect of DF on observed means in their samples.

With nonuniform DF, some focal group respondents will have higher ESs than the reference group respondents with equal theta scores, while the reverse will be true of other focal group members. Due to their signed nature, such differences are allowed to cancel across respondents such that sizable item DF could be present, yet the item SIDS as a whole could equal zero due to the averaging of positive and negative ES differences. In sum, with respect to the taxometric framework dimensions, SIDS is an (a) item-level index that (b) uses actual focal group sample data, (c) allows cancellation of DF across respondents, and (d) is presented in the metric of the observed scores.

A similar index that takes the absolute value of this ES difference for each respondent prior to averaging can be computed from

$$UIDS_i = \frac{\sum_{s=1}^{N} \left| ES_{(si|\hat{\theta},\gamma_F)} - ES_{(si|\hat{\theta},\gamma_R)} \right|}{N} \tag{4}$$

where UIDS_i represents the unsigned item difference in the sample for item i. This index is conceptually similar both to Raju's NCDIF index and to a polytomous extension of Camilli and Shepard's (1994) UPD-θ index. However, those indices square the difference prior to averaging whereas the UIDS statistic uses absolute value to retain the natural metric of the ESs. In effect, the UIDS represents the hypothetical amount of DF present in the item had the DF been uniformly against one (either the focal or the reference) group. Due to the absolute value in the UIDS formula, there is no canceling of DF across levels of theta. Thus, with respect to the taxometric framework, UIDS (a) is item-level, (b) uses sample data, (c) does not allow cancellation across respondents, and (d) is in the observed score metric.

One benefit of reporting both SIDS and UIDS is that the researcher can get an indication of the amount of cancellation of DF that takes place across respondents. If SIDS and UIDS are equal, the DF present is uniform in nature such that one group always has higher ESs than the other. If, however, the UIDS is larger than SIDS, cancellation of DF occurs across levels of theta, and it is strongly recommended that the researcher examine the ES plots to better understand the nature of the item DF.

Both SIDS and UIDS are based on estimated theta scores of the actual sample. As such, these statistics are said to describe the amount of DF encountered by focal group respondents in the current sample. These estimates of DF will generalize to the broader population to the extent that the current sample was randomly sampled from the broader population of interest.

D-Max

It may be useful to know the maximum extent to which any one person in the focal group is affected by item DF. D-Max can be reported as the maximum SIDS (either positive or negative) in the sample. D-Max is reported as a signed index, so that the researcher may know whether the focal group member is at an advantage or disadvantage due to item DF present. The interpretation of D-Max is as the largest difference in expected response for any one respondent in the sample. With respect to the taxometric framework, D-Max (a) is item-level, (b) is based on sample respondents, (c) does not consider cancellation, and (d) is presented in the metric of observed scores. D-Max is not intended to be a direct indicator of item or scale quality. Rather, it indexes the interplay between the DF present and the specific characteristics of the sample.

Expected Score Version of Cohen's *d*

An additional measure of item DF can be computed using an ES version of Cohen's (1988) *d*. With this index, ESs are computed for the focal group using both sets of item parameters as described earlier. Then the mean and pooled standard deviation are computed and an ES standardized difference (ESSD) is computed as

$$ESSD_i = \frac{\overline{ES}_{(\gamma_F)} - \overline{ES}_{(\gamma_R)}}{SD_{ItemPooled}} \tag{5}$$

$SD_{\mathrm{ItemPooled}}$ is computed as

$$SD_{ItemPooled} = \sqrt{\frac{(N_F - 1)SD_{ES(i|\gamma_F)} + (N_F - 1)SD_{ES(i|\gamma_R)}}{2 * N_F - 2}} \tag{6}$$

where N_F is the sample size in the focal group. As ESSD is expressed in standard deviation units, it can be interpreted using Cohen's (1988) guidelines for small, medium, and large effect sizes. ESSD is similar to an item-level version of Stark et al.'s (2004) test-level d_{DTF} index (described later). With respect to the taxometric framework, ESSD (a) is item level, (b) is based on focal group sample data, (c) allows cancellation across respondents, and (d) is presented in the metric of standard deviations.

A set of indices that are based on an assumed normal distribution rather than sample data may also be developed. The indices are very similar to those described above, aside from the assumption of normality of the focal group latent trait distribution. Their development is detailed in Appendix B. Table 1 provides a summary of item-level indices and their interpretations.

Scale-Level Indices

From these item-level DF effect size statistics, a number of scale-level statistics can be developed. Two indices are the sums of the signed and unsigned differences across the items,

$$STDS = \sum_{i=1}^{j} SIDS_i \tag{7}$$

$$UTDS = \sum_{i=1}^{j} UIDS_i, \tag{8}$$

where STDS and UTDS represent the signed and unsigned test difference in the sample. The term *test* rather than *scale* is used to provide consistency with previous indices and work on this topic (e.g., Flowers et al., 1999; Raju et al., 1995, 2009) and acronym clarity.

Conceptually, differences in ESs are taken for each participant (just as with SIDS) for each item. These are summed across the *j* items for each respondent and then averaged across the *N* respondents. This reduces mathematically to the sum of the SIDS measure (which averages across respondents prior to items). With respect to the taxometric framework, both STDS and UTDS (a) are scale level, (b) use focal group respondent data, and (c) are in the metric of observed scores. The STDS indices allows for full cancellation of DF across both respondents and items. Conversely, the UTDS allows no cancellation across either persons or items.

STDS has a clear interpretation as the averaged, across focal group respondents, difference in expected scale scores. An advantage of the STDS, as for its item-level counterpart, is that it retains the metric of the summed scale score. Thus, an STDS of −2.5 would indicate that on average, focal group respondents would be expected to score 2.5 units lower on the summed scale than would reference group respon-

Table 1
Summary of Item-Level DF Effect Size Indices and Interpretations

Item-level index	Notes and interpretations
Signed item difference in sample (SIDS)	The average difference in ESs across focal group sample respondents. DF across respondents is allowed to cancel in cases of nonuniform differences in ESs.
Unsigned item difference in sample (UIDS)	The average difference in ESs across focal group sample respondents had differences been uniform in nature (i.e., always favoring one group). Comparing UIDS and SIDS gives an indication of the extent to which differences in ESs cancel across different respondents. If SIDS and UIDS differ, examination of ES plots is strongly recommended.
Signed item difference in normal distribution (SIDN)	The average difference in ESs across a normal theta distribution. DF is allowed to cancel across regions of theta in cases of nonuniform differences in ESs.
Unsigned item difference in normal distribution (UIDN)	The value of average difference in ESs across a normal theta distribution, had differences been uniform in nature (i.e., always favoring one group).
Maximum difference in sample (D-Max)	The maximum difference in ESs for any member of the focal group sample. The theta value of the respondent who experiences D-Max is also reported by the VisualDF program.
Expected score standardized difference (ESSD)	An ES version of Cohen's *d*. Mean ESs are computed for the focal group sample respondents using both focal and reference item parameters. The difference between these means are divided by the pooled *SD* of the two sets of ESs. The metric can be interpreted using the guidelines given by Cohen (1988).

Note. In all cases, focal group expected scores are computed using item parameters estimated in both the focal group sample and the reference group sample. Nonuniform differences in expected scores indicate that for some respondents (i.e., some areas of theta), expected scores will be higher for the focal group and for other respondents (i.e., other areas of theta) expected scores will be higher for the reference group. DF = differential functioning; ES = expected score.

dents with equal standing on the latent trait. This difference would be expected solely due to the DF present in the scale.

When interpreting the magnitude of STDS, it is helpful to report the possible scale score range. For instance, if there are six items with five response options each, the range of the scale score is from 6 to 30. Conversely, SIDS can also be averaged across items rather than summed. This may be useful when comparing scales with differing numbers of items. Note, however, that averaging is not recommended when items within a given scale have differing numbers of response options. Items with different numbers of response options will have different potential ranges, obfuscating the interpretation of the average. Standardized indices are a better choice for such comparisons.

As with examining item-level DF effect size measures, examining STDS alongside UTDS gives an indication of the amount of cancellation of DF that takes place across both items and respondents. Large differences in the STDS and UTDS statistics indicate that much cancellation takes place among items and/or respondents, whereas similar STDS and UTDS indices indicate that most DF is relatively unidirectional in nature.

Another potential index is one in which ETSs are computed and the absolute values of differences in respondents ETSs are averaged,

$$UETSDS = \frac{\sum_{s=1}^{N}[|ETS_{(s|\hat{\theta},\gamma_F)} - ETS_{(s|\hat{\theta},\gamma_R)}|]}{N} \tag{9}$$

where UETSDS is the unsigned expected test score difference in the sample. As the sum of the item ESs, ETSs allow cancellation of DF across items for each respondent *s*. However, cancellation is not allowed across respondents due to the absolute value of the difference that is taken. The UETSDS index bears a strong conceptual resemblance to Flowers et al.'s (1999) DTF index (which is a polytomous extension of Raju et al.'s 1995 DTF), in that DTF is equivalent to the squared UETSDS. UETSDS can be described as (a) scale level, (b) sample based, (c) cancellation of DF across items but not across persons that is (d) in the metric of observed scores. Note that the average, across respondents, of differences in ETSs without an absolute value is equal to the STDS.

The UETSDS provides a compromise with respect to cancellation of DF, allowing cancellation across items but not persons. Allowing cancellation across items is inherently appealing because it reflects the true behavior of DF on observed scale scores. The interpretation of the UETSDS is that it reflects the hypothetical difference in expected scale scores that would have been present if scale-level DF had been uniform across respondents (i.e., always favoring the focal group).

Also available is a test-level ES version of Cohen's *d*,

$$ETSSD = \frac{\overline{ETS}_{(\gamma F)} - \overline{ETS}_{(\gamma R)}}{SD_{TestPooled}} \tag{10}$$

where ETSSD represents the expected test score standardized difference and ETS is the expected test (scale) score. The ETSSD is very similar to Stark et al.'s (2004) d_{DTF} index. However, Stark et al. use an assumed normal distribution to compute ETSs for the numerator of the formula, and the ETSSD uses the ETS values of the actual sample respondents. Also, the Stark et al. (2004) index uses the focal groups' observed scores to compute the denominator of the equation, whereas ETSSD uses the pooled standard deviation of the estimated ETSs. The ETSSD (a) is scale level, (b) is based on observed scores, (c) allows cancellation of DF across both items and respondents, and (d) is presented in the metric of standard scores.

Scale-level D-Max (called Test D-Max) is also readily computed from the ETSs. Test D-Max (a) is scale level, (b) is based on sample data, (c) allows cancellation across items but does not consider cancellation across respondents, and (d) is in observed score metric.

Another potential index of interest is one that indicates the range of theta over which the focal group is at a disadvantage.[4] This index, called region of disadvantage, is determined by the range of

[4] Special thanks to Alan Mead for suggesting this index.

the sample estimated theta scores for which ETSs are lower when computed using focal group parameters than reference group parameters. As ETSs can take on complex nonlinear forms, the region of disadvantage need not be continuous.

As with item-level indices, scale-level indices may be based on an assumed normal distribution rather than sample theta-estimates. Appendix B describes normal-distribution-based scale-level indices, and Table 2 provides a summary of the scale-level indices proposed.

In sum, the taxonomy presented here can be used to classify DF effect size indices with respect to four dimensions: (a) item- or scale-level indices, (b) use of either sample focal group data or an assumed focal group distribution, (c) the extent of allowance for cancellation of DF across respondents and across items, and (d) whether the index is in the metric of observed scores or is standardized in some way. A summary of how the newly developed and existing DF effect size indices fit into the taxometric framework can be found in Tables 3 and 4. As it makes little sense to allow cancellation of DF across persons but not items, no scale-level indices are proposed for those cells.

Cautions on the Use of Scale-Level Indices

The STDS, ETSSD, and Stark et al.'s (2004) DTFR and d_{DTF} indices allow for cancellation of DF both across items and across persons (or levels of theta). In effect, they indicate the effect of DF on scale scores, on average, across theta values. Very often this may not be the primary information of interest to the researcher. For example, imagine an employee selection context in which hiring decisions are made around a cutoff score that is near the top of the potential range of scores. In this case, DF at low levels of theta matters little, as those respondents are not near the cutoff value. If only scale scores are used, it is appropriate to allow item-level cancellation across items, within respondents. However, cancellation across levels of theta may still be highly undesirable, as there may be some subset of respondents for which large DF is present. Stark et al. (2004) acknowledged this limitation of their DTFR and d_{DTF} indices and developed an alternative index to investigate differences in selection ratios that may result at a given theta level.

Although the indices discussed here provide information related to the extent of DF that exists, the examination of plots of ESs along the theta distribution is often as informative. Such an examination is possible both for individual items and for ETSs. Plots allow the researcher to determine the areas of theta at which DF is encountered at both the item and test levels. Examining these plots alongside a sample theta frequency distribution provides a visual indicator of why there may be differences in SIDS and SIDN. For example, if ESs differ greatly at very low levels of theta, this may affect SIDN but not SIDS if there are no respondents with very low theta levels in the sample.

Software

Measures of DF effect size are unlikely to be widely used unless they are readily obtained. Moreover, the computations behind these indices are cumbersome though not complicated. A Microsoft Excel-based computer program called VisualDF was developed in order to compute all of the effect sizes presented. The software, coded using MS Excel's Visual Basic for Applications, is free and available for anyone to modify or use; it is available for download from Adam W. Meade's website. (The URL is http://www4.ncsu.edu/~awmeade)

Table 2
Summary of Test-Level Effect Size Statistics and Interpretations

Test-level index	Notes and interpretations
Signed test difference in the sample (STDS)	The difference in observed summed scale scores expected, on average, across all focal group sample respondents, due to DF alone. Allows cancellation of DF across both items and persons.
Unsigned test difference in the sample (UTDS)	The expected difference in observed scale scores averaged across all focal group sample respondents that would have been exhibited had all DF been uniform in nature. Comparing UTDS and STDS provides an index of the extent to which DF in items tends to be nonuniform across items and/or theta levels. No cancellation of DF is allowed across items or persons.
Stark's DTFR	The difference in observed summed scale scores expected, on average, across a hypothetical focal group with a normally distributed theta, due to DF alone. Allows cancellation of DF across both items and persons.
Unsigned DTFR (UDTFR)	The difference in observed summed scale scores expected, on average, across a hypothetical focal group with a normally distributed theta, had DF been uniform in nature for all items.
Unsigned expected test score difference in the sample (UETSDS)	The hypothetical difference in expected scale scores that would have been present if scale-level DF had been uniform across respondents (i.e., always favoring the focal group). Allows cancellation of DF across items but not persons.
Unsigned expected test score difference in normal distribution (UETSDN)	Identical to UETSDS but computed using a normal distribution.
Expected test score D-Max (Test D-Max)	The value of the largest ETS difference for any one member of the sample focal group. In other words, the largest value of DF experienced by any one respondent in the sample.
Region of disadvantage	Area of theta in which the focal group expected scale scores computed using focal group parameters are lower than those computed using reference group parameters.
Expected test score standardized difference (ETSSD)	An ETS version of Cohen's *d*. The metric can be interpreted using the guidelines on effect size given by Cohen (1988).

Note. In all cases, focal group expected scores are computed using item parameters estimated in both the focal group sample and the reference group sample. Nonuniform differences in expected scores indicate that for some respondents (i.e., some areas of theta), expected scores will be higher for the focal group and for other respondents (i.e., other areas of theta) expected scores will be higher for the reference group. DF = differential functioning; ETS = expected test score.

Table 3
Taxonomy of Item- and Scale-Level DF Indices

	DF cancels across items?			
	Yes DF cancels across respondents/theta?		No DF cancels across respondents/theta?	
Theta used	Yes	No	Yes	No
Sample	STDS ETSSD[a]	UETSDS Test D-Max Region of disadvantage Flowers et al.'s (1999) DTF	*SIDS* *ESSD*[a]	*UIDS* *Test D-Max* UTDS
Assumed distribution	Stark's DTFR Stark's d_{DTF}[a]	UETDSN	*SIDN*	*UIDN* UDTFR

Note. Italics indicates an item-level index; normal font indicates a scale-level index. DF = differential functioning; STDS = signed test difference in the sample; UETSDS = unsigned expected test score difference in the sample; SIDS = signed item difference in sample; UIDS = unsigned item difference in sample; Test D-Max = maximum difference in expected test score for sample; ETSSD = expected test score standardized difference; DTF = differential tax functioning; ESSD = expected score standardized difference; UTDS = unsigned test difference in the sample; Stark's DTFR = Stark et al.'s (2004) DTFR; UETSDN = unsigned expected test score difference in normal distribution; SIDN = signed item difference in normal distribution; UIDN = unsigned item difference in normal distribution; Stark's d_{DTF} = Stark et al.'s (2004) d_{DTF}; UTDFR = unsigned DTFR.
[a] Indicates that index is standardized; other indices are in the metric of observed scores.

Two Examples

Two examples were chosen from different research domains to illustrate the proposed effect sizes and accompanying software. In one domain, large and pervasive DF is often encountered, and in the other very small levels of DF are typically found.

Cross-Cultural Example

Invariance investigations of measures that have been translated and used in a different culture often show large and significant amounts of DF (for a recent review of 130 such studies, see Chen, 2008). Data for the first illustration were taken from a DF study of the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) conducted by Baranik et al. (2008). Comparisons between U.S. respondents and Taiwanese respondents answering items in Chinese were examined with the LRT. Some item parameters had to be collapsed across categories due to a lack of use of some item response options, and one item was excluded from analyses due to estimation problems (for details, see Baranik et al., 2008).

The sample size was 334 in the Chinese-speaking Taiwanese focal group and 464 in the English-speaking U.S. reference group. Statistical significance of item DF was determined using IRTLRDIF (Thissen, 2001), and six of the nine items analyzed showed statistically significant DF across the two samples (see Baranik et al., 2008). Item parameters and focal group data were then input into VisualDF, which was used to compute expected a posteriori theta estimates for focal group members, DF effect size measures, and ES plots.

Cross-Administration Format Example

The second comparison was pulled from a study by Meade, Michels, and Lautenschlager (2007) in which the invariance of several personality scales was compared across administration formats. Several studies have found relatively small differences between Internet-administered and paper-and-pencil-administered measures, in contrast to cross-cultural research with language differences (Bartram & Brown, 2004; Ployhart, Weekley, Holtz, & Kemp, 2003; Potosky & Bobko, 2004; Rosenfeld, Doherty, Vicino, Kantor, & Greaves, 1989; Whitaker & McKinney, 2007). Respondents completed several subscales of the OPQ32 (SHL, 2000), and from these, the six-item Outgoing scale was chosen for comparison. In total, 373 respondents completed the online version of the Outgoing scale and 375 completed a paper-and-pencil version of the same measure. Meade, Michels, and Lautensclager used a confirmatory factor analytic approach to assess invariance; data from the Outgoing scale were reanalyzed using IRTLRDIF and VisualDF.

Table 4
Estimated Item Parameters and G^2 Statistics for Cross-Cultural RSES Comparison

			U.S. sample				Taiwanese sample			
Item	G^2	*df*	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3
1	1.6	2	2.2	−0.05			1.73	−0.05		
2	90.4*	2	2.11	−0.13			1.79	0.88		
3	0.3	2	2.74	0.11			2.5	0.13		
4	10.5*	2	1.68	0.45			2.04	0.09		
5	54.5*	2	2.64	0.27			1.04	1.51		
6	113.6*	3	2.37	−1.04	0.75		0.6	−1.63	0.8	
7	65.6*	3	1.81	−0.95	1.23		2.32	−0.84	0.26	
8	84.5*	4	1.58	−1.81	0.11	1.62	2.37	−1.67	−0.45	0.34
9	6.2	3	1.59	−0.47	0.79		1.97	−0.22	0.72	

Note. a, b_1, b_2, and b_3 are sample item parameters. RSES = Rosenberg Self-Esteem Scale (Rosenberg, 1965).
* $p < .01$.

Results

Cross-Cultural Item-Level Comparison

Estimated item parameters and associated G^2 statistics for the U.S. versus Chinese-speaking Taiwanese samples are given in Table 4, and corresponding item-level DF effect size indices are given in Table 5. As shown in Table 4, six of the nine items were identified as DF by the LRT. ES plots in Figure 2 provides a visual representation of DF.[5] Several aspects of the investigation are notable. First, although G^2 is effective at identifying statistically significant DF, it is not a good indicator of the amount of DF present in any given item. For example, Item 6 has the largest G^2 value, yet the DF present in the item did not result in the largest difference in ESs. Examination of Figure 2 indicates that Item 6 showed a large amount of cancellation of DF in the sample. This cancellation is also apparent when one compares the SIDS and UIDS (and SIDN and UIDN).

Items 1, 3, and 9 did not show statistically significant DF, and the effect size statistics and ES plots confirm the low magnitude of DF present in those items. For instance, Item 3 showed similar item parameters across groups, a G^2 value less than 1.0 and a SIDS of −.003. This indicates that across all focal group sample respondents, the average expected difference in response for persons with equal theta values in the focal and reference groups would be .003. The UIDS indicates that even had the DF been uniform in nature (i.e., always favoring one group), the effect size would still have been only a .013 difference (on a 2-point scale) between the two groups.[6]

Examination of Table 5 indicates that the largest DF found (D-Max) for any Taiwanese respondent on any item was .749 (Item 8). After categories were collapsed, this item had the largest number of response options (4). Thus, for some respondents in the Taiwanese sample, one would expect the response to be .749 (on scale of a 1–4) higher than for someone in the U.S. sample with a corresponding theta score.

Item 2 showed identical absolute values of SIDS and UIDS and between SIDN and UIDN, indicating uniform DF across all respondents (Figure 2 confirms this). Although the other statistics for this item are lower than those for Item 8, Item 2 has only two response options (after collapsing categories) whereas Item 8 has four. As a result, most of the effect size indices are interpreted on a different metric for the two items. One exception is the ESSD statistic, which is directly comparable across items with different numbers of response options, as the metric is given in standard deviation units. ESSD is strongly advised for the comparison of DF of items with different numbers of response options.

Response Format Item-Level Comparison

Table 6 contains estimated item parameters for the online versus paper-and-pencil example as well as G^2 statistics. Table 7 contains item-level DF effect size indices, and ES plots for the six items are given in Figure 3. As shown in Table 6, no items exhibited statistically significant DF across groups. Comparison of Tables 5 and 7 reveals that effect size statistics are typically lower for the administration format example. Direct comparisons across these two examples are confounded by the fact that effect sizes are reported in the metric of the item response scale of each item. For the format example, there were sufficient data to make use of all five response options for all items. However, many of the items in the cross-cultural comparison were collapsed into a much smaller number of response options. Thus, the effect size statistics have a different potential range. Even so, the small values of the DF effect size statistics in Table 7 are quite small, with no item showing DF greater than .15 (on a 1–5 scale). Item 4 exhibited the most DF, and even then the effect was not large. An interpretation of the SIDS would be that respondents using the online format would be expected to have observed scores .114 lower (on a 5-point scale), on average, than their paper and pencil counterparts with identical levels of the outgoingness trait. Examination of ES plots reveals that that the largest difference appears to be for Item 5. This difference is also revealed by the D-Max statistic in Table 7. Item 5 has very low SIDS and SIDN values, however, due to the canceling effects of DF across levels of theta.

ES standardized differences are directly comparable across both items and example data sets. As can be seen, even the largest ESSD in the online example (Item 4) is smaller than that of all of the significantly DF items in the cross-cultural example.

Scale Score DF

Table 8 contains scale-level DF statistics for both examples, and Figure 4 contains ETS graphs. Examination of ETSs in Figure 4 reveals relatively minor differences between the ETS in both example data sets. Although this is fully expected, given the administration format example results, this graph makes explicit the notion of cancellation of DF (see Flowers et al., 1999; Raju et al., 1995, 2009) for the cross-cultural example. As is evident here, one limitation of ETS plots is that ETS lines will appear to be very similar, even when DF is present, when the range of potential scale scores is very large. It is for this reason that plots are not always sufficient indicators of DF.

Examination of test-level statistics in Table 8 highlights the differences between ETSs. As can be seen, the DF effect size was somewhat larger in the cross-cultural comparison, even with the cancellation of DF that occurred across items. The extent of cancellation of DF was most obvious when observing the sizable difference in the STDS and UTDS in the cross-cultural example. The UTDS indicates that had the DF been uniform, the Taiwanese group would have been expected to score 1.518 observed scale score units higher on average than a U.S. sample with identical levels of theta. The 1.518 is directly interpretable given the scale's

[5] Note that VisualDF can provide smooth line plots using sample data or plots in which each respondent is indicated with a small marker. The latter can useful for illustrating whether or not differences in ES at a given level of theta may be practically important (e.g., if very few respondents have theta values of that level). However, with large samples, markers for individual respondents can be cumbersome. For this reason, smooth plots are presented here.

[6] Interpretation of probabilities of response with collapsed response options is more cumbersome. For instance, when five response options have been collapsed into two, ESs indicate the probability of responding with a 1–4 versus responding with a 5 (though the ES metric is between 1 and 2 rather than the traditional probability 0–1 metric). It is the difference in these ESs across groups that is indexed with the effect size measures. Although such interpretations are inherently less straightforward than those of uncollapsed data, collapsing categories is common, and the example provided here provides a realistic view of such instances. Standardized indices, such as ESSD, are strongly recommended for comparing DF of items with different numbers of response options.

Table 5
Item-Level Effect Size Statistics for Cross-Cultural RSES Comparison

Item	SIDS	UIDS	SIDN	UIDN	D-Max	ESSD
1	0.008	0.037	0.005	0.036	−0.054	0.028
2	−0.273	0.273	−0.259	0.259	−0.456	−1.087
3	−0.003	0.013	−0.004	0.013	−0.029	−0.008
4	0.091	0.091	0.088	0.088	0.177	0.350
5	−0.148	0.167	−0.161	0.180	−0.504	−0.640
6	0.088	0.192	0.083	0.215	0.508	0.261
7	0.193	0.219	0.192	0.222	0.482	0.429
8	0.417	0.424	0.402	0.420	0.749	0.726
9	−0.085	0.101	−0.073	0.094	−0.158	−0.184

Note. RSES = Rosenberg Self-Esteem Scale (Rosenberg, 1965); SIDS = signed item difference in sample; UIDS = unsigned item difference in sample; SIDN = signed item difference in normal distribution; UIDN = unsigned item difference in normal distribution; D-Max = maximum difference in sample; ESSD = expected score standardized difference.

range from 9 to 23. Further information is garnered by examining the UETSDS, which allows cancellation of DF across items (as would naturally occur) but not persons. Given a STDS of .288, a UETSDS of .295, and a UTDS of 1.518, it is clear that much of the DF present cancels across items rather than across respondents. In other words, it would seem that a substantial amount of the DF present tends to be uniform within items but that different items favor different groups. This DF is allowed to cancel out at the scale level, resulting in similar UETSDS and STDS values.

An interesting analysis is the comparison of differences between observed scores and ETSs. The observed mean on the RSES was 14.02 for the U.S. sample and 13.36 for the Taiwanese sample (a difference of −.66). However, the STDS for these groups was .29, indicating that, had latent trait scores been identical across the two samples, the Taiwanese group would have had mean ESs that were higher by .29 points. We can state, by extension, that had there been no DF on the RSES, U.S. respondents would have scored approximately .95 points higher than the Taiwanese respondents. The DF that was present served to decrease group differences in observed scores in this case.

Examination of the ETSSD and observed score Cohen's d tells a similar story. For the cross-cultural example, the ETSSD indicates that the standardized difference expected between the group means would be .165 standard deviations higher in the Taiwanese group. However, the observed score d was −.215, indicating that the Taiwanese group had a lower mean than the U.S. group. The discrepancy between the numbers is attributable to differences in the latent means among the two groups. Without DF, U.S. sample means would have been even higher than those of the Taiwanese group.[7] By providing DF effect sizes in the metric of observed scores, researchers can tease apart apparent differences in group means due to DF and those due to differences in latent traits. In effect, this allows researchers to "correct" group mean differences for DF and indicate what the difference in group means would have been had DF not been present.

Discussion

Suggestions for Use of Effect Size Measures

Guidance is provided with respect to how a researcher might use the program given his or her particular needs. In addition, a goal of developing DF effect size measures is that eventually meta-analyses can be conducted related to the amount of DF present under certain conditions. As such, some indices are universally recommended to allow for future aggregation of DF evidence.

Of the proposed indices in the taxometric framework, the STDS, UETSDS, and ETSSD should be reported regardless of the data analytic situations. These indices will provide essential information as to the amount of DF present in the sample and the extent to which this DF cancels across respondents (by comparing STDS and UETSDS). Moreover, the ETSSD provides a normed statistic useful in comparisons across studies with different numbers of response options and items. Other indices may be reported as needed, based on the particular needs of the researcher. There are many reasons why a researcher may examine DF, and these reasons have a direct bearing on the type of effect size that is needed. Prior to analyzing any data, the researcher should answer the following questions.

1. Is item-level DF of interest, or is scale-level DF the focus? If all comparisons are to be made at the scale level and the researcher is comfortable with the notion of cancellation of DF across items, scale-level effect size statistics should be the focus of the effect size investigation. In such cases, item-level DF can be examined but should not be of primary concern.[8]

[7] The ETSSD and observed d are not identical as the ETSSD uses ETSs for the focal group and uses both sets of item parameters; thus, the sample size of the two sets of ETSs is the same. Observed d uses the actual reference group observed mean and variance in computations. The ETS variance computed using the focal group only and the observed variance for the reference group may differ substantially. For this reason, it is imprecise to estimate the value of d had DF not been present when using ETSSD. It would be more accurate to use STDS to alter the numerator of the d formula and to leave the denominator unchanged.

[8] However, this issue is not as straightforward as it may seem. If there are severe differences in the measurement properties of a scale across groups, it is possible that the construct validity of the instrument is compromised in such a way that comparisons simply are not valid. If items do not relate to each other in similar ways, the construct itself is defined differently in the two groups and any comparison between groups is specious (Vandenberg & Lance, 2000). In an IRT framework, this would be manifest as radically different item a parameters across the two groups, as a parameters indicate the strength of relationship between the item and the latent construct. However, if the overall construct seems comparable across groups (and a factor analysis results in a common structure and similar factor loadings), the researcher may move on to an investigation of DF and DF effect size.

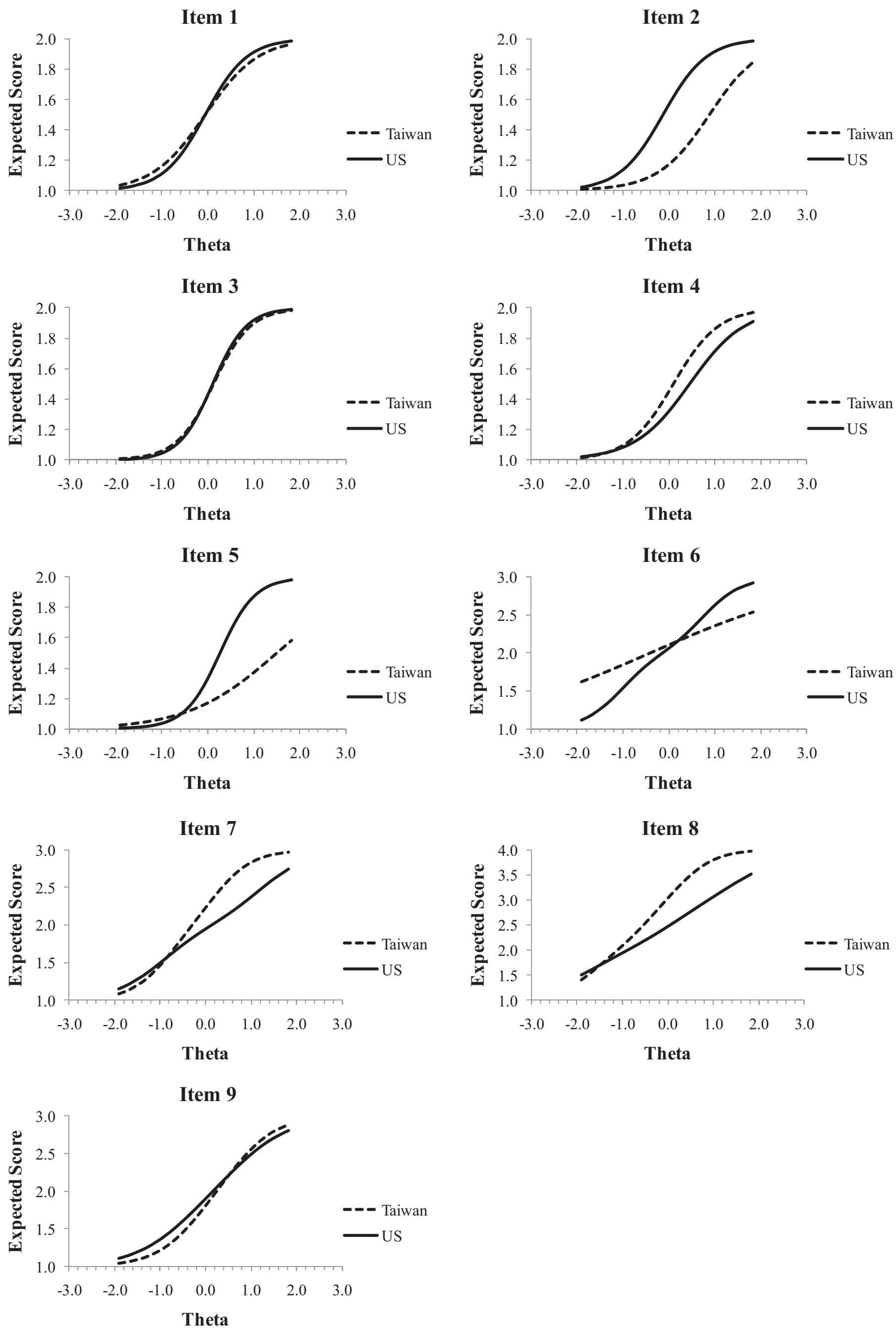

Figure 2. Item expected score plots for cross-cultural RSES comparison. RSES = Rosenberg Self-Esteem Scale.

2. Is the sample used that of interest? Some indices indicate the amount of DF encountered by the current sample. If the researcher is interested in the DF encountered by this sample in particular, these are the statistics that should be used. For example, a researcher wishing to compare means on a leadership effectiveness survey across business units may be interested in his or her current sample. The SIDS and STDS indices can allow a researcher to determine the extent to which differences seen are due to DF for these respondents. In other cases, the population DF is of interest. IRT parameters are invariant across subpopulations, and estimates obtained in different samples differ only by a linear transformation (Hambleton & Swaminathan, 1985), making the use of nonrandom samples possible. For example, in examining a personality selection test for DF across gender groups, one's focus is on determining how much DF will be present for the population of all applicants. If the sample is large, randomly selected, and

Table 6
Estimated Item Parameters and G^2 Statistics for Format Administration Comparison

			Paper-and-pencil sample					Internet sample				
Item	G^2	*df*	*a*	b_1	b_2	b_3	b_4	*a*	b_1	b_2	b_3	b_4
1	1.0	5	1.21	−2.13	−1.66	0.02	0.55	1.19	−2.17	−1.68	0.14	0.59
2	1.3	5	1.98	−1.04	−0.96	−0.32	0.05	1.9	−1.15	−1.02	−0.33	0.02
3	3.1	5	1.28	−1.55	−1.25	0.08	0.48	1.56	−1.51	−1.23	0.05	0.45
4	5.1	5	1.99	−1.43	−1.14	−0.38	0.23	1.97	−1.34	−0.99	−0.23	0.22
5	10.7	5	1.93	−1.01	−0.88	−0.25	−0.12	2.72	−0.92	−0.75	−0.28	−0.07
6	9.0	5	2.78	−0.85	−0.71	−0.04	0.08	2.82	−0.97	−0.78	−0.11	0.15

Note. No G^2 values were significant at $p < .05$ level. a, b_1, b_2, b_3, b_4 are sample item parameters.

representative of the population, the SIDS and STDS are most appropriate. However, if the current sample is considered nonrandom and the population is assumed to be normally distributed, the SIDN and UIDN statistics may provide the most beneficial information.

3. Is the focus on DF on the average or at all levels of theta? In some instances, only group mean comparisons are of interest to the researcher (e.g., comparing average job satisfaction ratings across U.S. and Korean workers). In contrast, consider a situation in which DF across racial groups for a job selection test is assessed. In this case, DF around the test cutoff score is of much more interest than is DF at the mean level of test scores (presuming the cutoff is not at the mean). In such situations, investigating the difference in expected test or item scores at a particular theta level is the most appropriate approach. In general, if only group means will be compared, cancellation of DF across theta is appropriate, and STDS, SIDS, ETSSD, ESSD, and Stark et al.'s (2004) DTFR and d_{DTF} are the primary statistics of interest. However, if individual respondents will be compared across different groups (e.g., for selection or performance appraisal purposes), UETSDS, UETDSN, UTDS, and UIDS should be examined and minimized to the extent possible.

4. How much does DF matter? Sometimes DF is extremely important, and sometimes it is less so. Consider the case in which a researcher wants to know if facets of job satisfaction have similar correlations across age-groups. In this case, if DF exists in similar degree across the facet measures, the correlation may be robust to DF even though the group means may differ. Conversely, researchers may wish to be extremely conservative with respect to DF across racial groups in selection tests. In such cases, even minor DF may result in a decreased probability of hire for a minority group. For such instances, even though scale-level DF may be of primary interest, items may be inspected for DF in order to determine the effect of each item on scale-level DF.

Summary of Indices

STDS. The STDS gives the actual difference in ETSs in the metric of observed scale scores, averaged across the focal group respondents. Thus, STDS has perhaps the most straightforward interpretation of any index. The STDS indicates the difference in observed score means that is expected due to DF alone. As shown in the examples, this information can be used directly in conjunction with observed means to determine approximately what the observed score difference between samples would have been had DF not been present. Of the indices in the taxonomy, this index will be the focal index in most instances. With respect to the earlier questions, this index is most useful when (a) scale scores are of primary interest, (b) the current sample respondents are the primary focus (or the sample is considered representative of the population of interest), and (c) the focus is on DF on the average across samples. Examples might include comparing leadership performance ratings across rater groups (e.g., peers, supervisors, subordinates) or examining organizational climate results across cultures. The STDS would allow researchers to determine what mean differences would have been had DF not been present.

UETSDS. The UETSDS is also a useful index in that it indicates the amount of DF present at the scale level had the ETSs uniformly favored one group. Item-level DF is canceled out by first summing ESs for each respondent. However, prior

Table 7
Item-Level Effect Size Statistics for Format Administration Comparison

Item	SIDS	UIDS	SIDN	UIDN	D-Max	ESSD
1	−0.036	0.036	−0.031	0.033	−0.049	−0.075
2	0.050	0.051	0.046	0.048	0.110	0.064
3	0.036	0.055	0.030	0.074	−0.135	0.061
4	−0.114	0.114	−0.097	0.097	−0.158	−0.158
5	−0.003	0.151	−0.021	0.173	−0.344	−0.004
6	0.071	0.071	0.056	0.056	0.133	0.074

Note. SIDS = signed item difference in sample; UIDS = unsigned item difference in sample; SIDN = signed item difference in normal distribution; UIDN = unsigned item difference in normal distribution; D-Max = maximum difference in sample; ESSD = expected score standardized difference.

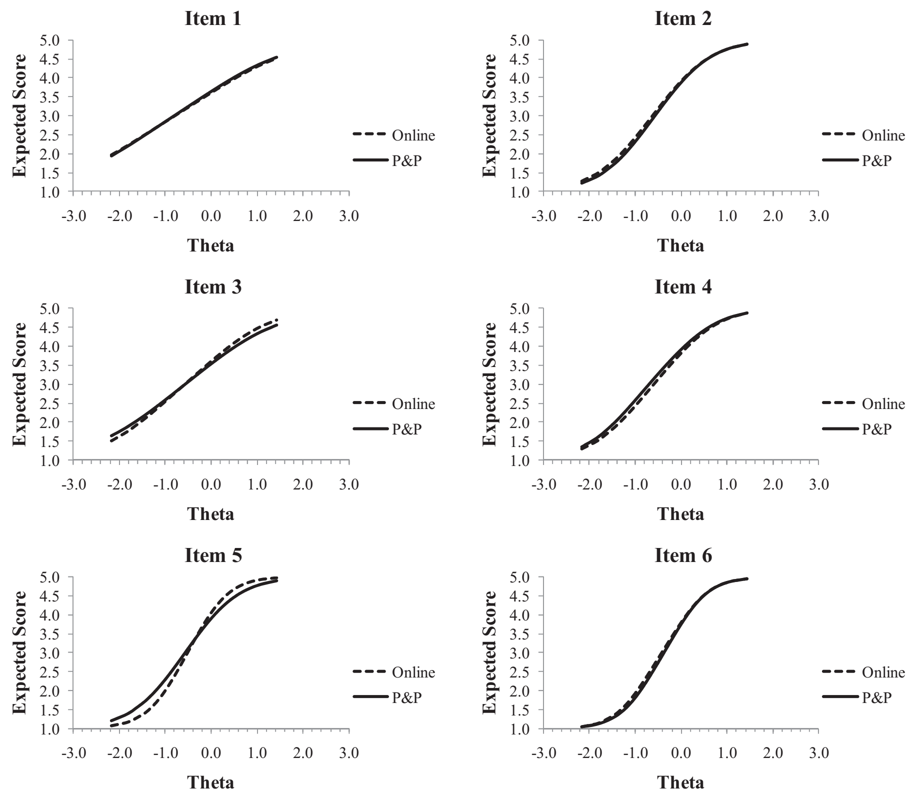

Figure 3. Item expected score plots for format administration comparison. P&P = pencil and paper.

to averaging ETSs, one takes the absolute value of the difference to prevent cancellation of DF across sample respondents. This index should always be reported. This index may be of special interest in contexts such as examining employee selection measures. For such measures, it may be permissible that DF is allowed to cancel across items (as only overall test scores are used in hiring), but it is not permissible that DF cancel across persons with different estimated theta values. Visual inspection of DF around the selection cutoff score would be important in this example.

Table 8
Test-Level Effect Size Indices

Test	Cross-cultural RSES example	Administration format example
STDS	0.288	0.004
UTDS	1.518	0.478
Stark's DTFR	0.273	−0.017
UDTFR	1.528	0.481
UETSDS	0.295	0.150
UETSDN	0.286	0.189
Test D-Max	0.358	−0.383
ETSSD	0.165	0.002
Theta of respondent with Test D-Max	0.078	−1.443
Theta region of disadvantage	1.51 to 1.82	−2.16 to −0.32
Potential scale range	9–23	6–30

Note. RSES = Rosenberg Self-Esteem Scale (Rosenberg, 1965); STDS = signed test difference in the sample; UTDS = unsigned test difference in the sample; Stark's DTFR = Stark et al.'s (2004) DTFR; UDTFR = unsigned DTFR; UETSDS = unsigned expected test score difference in sample; UETSDN = unsigned expected test score difference in normal distribution; Test D-Max = expected test score D-Max; ETSSD = expected test score standardized difference.

ETSSD and ESSD. These indices are, respectively, test and item latent score versions of Cohen's *d*. One advantage of ETSSD (and ESSD) is that the mean differences between groups is standardized on a metric for which Cohen's (1988) recommendations about small, medium, and large effect sizes can be applied. For this reason, ETSSD should always be reported. It is of primary use when (a) scale scores are of primary interest, (b) the current sample is of interest (or is considered a representative sample of the population of interest), (c) the focus is on average group differences rather than comparing individuals across groups, and (d) cancellation of DF across levels of theta is not of concern. ESSD can be used in similar circumstances for item-level comparisons. These indices will prove useful in cases of trying to compare measures with different numbers of response options (e.g., evaluating whether to report employee satisfaction survey results in a five-response option metric or a collapsed three-response option metric).

SIDS. If item-level DF is of interest, the SIDS provides the same basic information as STDS at the item level. Single item results are often reported in satisfaction and culture surveys. The SIDS index would allow researchers to adjust observed item scores in order to make more valid longitudinal comparisons of organizational support, for example.

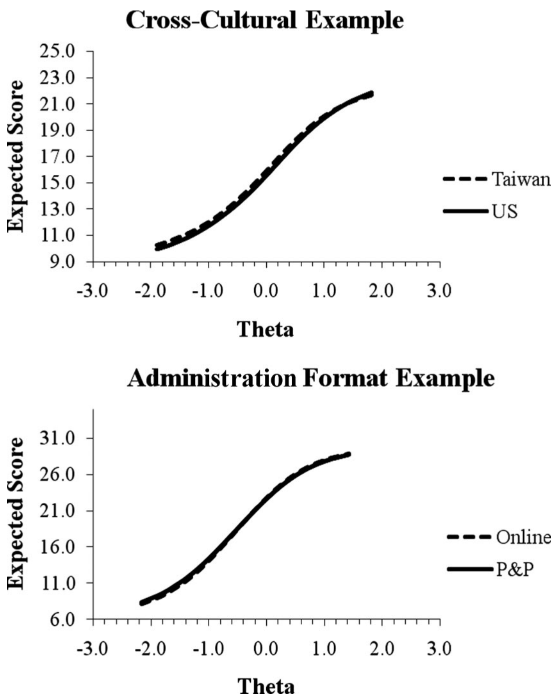

Figure 4. Expected test score plots for data examples. P&P = pencil and paper.

UTDS and UIDS. If the researcher is interested in the extent to which DF may be canceling across different areas of theta, comparing the UTDS and UIDS to the STDS and SIDS is informative. The UTDS is expressed in the metric of the observed scale scores and answers the hypothetical question "what would the difference in observed group mean scores due to DF have been if DF had been uniform in nature for all items?" UIDS also is expressed in the metric of the observed scores, but at the item level. It is difficult to imagine circumstances under which the UTDS and UIDS measures would be considered more central than the STDS and SIDS statistics. However, comparing the unsigned and signed measures can provide critical information about the nature of the DF. These comparisons would be critical in situations in which individual respondents may be compared on the basis of their observed scores, as in a selection context.

Stark's DTFR, UDTFR, UETSDN, SIDN, and UIDN. These indices assume a normal distribution for theta in calculations. It is possible that such indices would be preferred in cases in which the population is assumed to be normally distributed but the sample may not be representative (e.g., cases of examining DF of a selection measure across groups using a concurrent validation design using job incumbents). In this case, persons low on the construct being measured may have been screened out during the hiring process.

D-Max, Test D-Max, and region of disadvantage. It is rare that these indices would be the primary indices of interest. D-Max is useful for illustrating the full extent of DF that is present for any one respondent in the current sample. The region of disadvantage provides information related to where on the theta scale focal group members are at a disadvantage. These statistics complement other indices when DF cancels across some areas of theta.

Summary

Over the past two decades, significant progress has been made with methods of detecting statistically significant DF. However, a broader understanding and utilization of DF effect size is an essential next step in the progression of understanding invariance. Consistent reporting of the DF effect size indices presented here would allow accumulation of evidence for meta-analytic estimates of approximate DF effect sizes in different content domains. It is hoped that the availability of easy to use software in the form of the VisualDF program will facilitate this effort. Such estimates would give substantive (e.g., cross-cultural) researchers an idea of the extent to which DF may be expected in their particular domain of interest and would allow for a better understanding of conditions under which DF may or may not be expected.

References

Baker, F. B. (1995). Equate 2.1: Computer program for equating two metrics in item response theory [Software]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Baranik, L. E., Meade, A. W., Lakey, C. E., Lance, C. E., Hu, C., Hua, W., & Michalos, A. (2008). Examining the differential item functioning of the Rosenberg Self-Esteem Scale across eight countries. *Journal of Applied Social Psychology, 38,* 1867–1904.

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment, 12,* 278–284.

Bolt, D. M. (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.* Thousand Oaks, CA: Sage.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaption of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–353.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005–1018.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91–47). Princeton, NJ: Educational Testing Service.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86,* 215–227.

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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31,* 457–459.

Han, K. T., & Hambleton, R. K. (2007). *User's manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst: University of Massachusetts, School of Education.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Kim, S.-H. (2000, April). *An investigation of the likelihood ratio test, the Mantel test, and the generalized Mantel–Haenszel test of DIF.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15,* 269–278.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44,* 93–116.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31,* 430–455.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10,* 322–345.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata and situational judgment tests comparable? *Personnel Psychology, 56,* 733–752.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57,* 1003–1034.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33,* 133–147.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

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.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Rosenberg, M. (1965). *Society and the adolescent child.* Princeton, NJ: Princeton University Press.

Rosenfeld, P., Doherty, L. M., Vicino, S. M., Kantor, J., & Greaves, J. (1989). Attitude assessment in organizations: Testing three microcomputer-based survey systems. *Journal of General Psychology, 116,* 145–154.

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53,* 531–562.

Shealy, R., & Stout, W. (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,* 159–194.

SHL. (2000). *OPQ32 technical manual.* Boulder, CO: Author.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures.* Bowling Green, OH: Author.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497–508.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292–1306.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402–415.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Thissen, D. (2001). Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill: University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–172). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Erlbaum.

Whitaker, B. G., & McKinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods, 39,* 502–509.

Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the Satisfaction with Life Scale. *Personality and Individual Differences, 46,* 396–401.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

(*Appendices follow*)

Appendix A

Sensitivity of the LRT for Different Sized Samples

Simulation Method

Data for five polytomous items with five ordered response options were simulated using sample sizes of 100, 200, 500, or 1,000 per group. Item parameters were taken from the first five items of the reference group reported in Meade, Michels, and Lautenschlager (2007). Except for Item 4, item parameters were identical across the two groups (i.e., no DF). Item 4's lowest *b* parameter was changed such that it was .1 lower in Group 2 than Group 1, whereas Item 4's highest *b* parameter was changed such that it was .1 higher in Group 2 than Group 1. This pattern of DF is associated with slightly more central tendency in Group 2 than Group 1. It will cancel across levels of theta, such that the mean observed scores of the two groups would be identical, provided equal theta distributions (cf. Meade & Lautenschlager, 2004). Theta scores were drawn from a standard normal distribution in both groups. One hundred sample replications were simulated for each condition.

This level of DF was intended to be trivially small in magnitude. For reference, the population effect size for Item 4 was SIDN = −0.018 and UIDN = 0.019. The Stark's DTFR and the UDTFR for the entire five-item scale were the same as the SIDN and UIDN for Item 4 (as no other item exhibited DF). Also, the average SIDS and UIDS values across replications equal SIDN and UIDN. In other words, for an item with response options from 1 to 5, group means are expected to differ by less than .02.

Data were simulated with WinGen2 (Han, 2007; Han & Hambleton, 2007) and were analyzed with IRTLRDIF (Thissen, 2001). Only simultaneous tests of differences across all item parameters are presented here.

Table A1
Percentage of 100 Replication Found to Be Statistically Significant by the LRT

Item	N = 100	N = 200	N = 500	N = 1,000
1	19	4	7	5
2	0	0	1	1
3	2	0	0	2
4	**1**	**4**	**14**	**42**
5	0	4	0	1

Note. The boldface type in Row 4 indicates that this was the only item for which significant differences were simulated. LRT = likelihood ratio test.

Simulation Results

Given that there was no DF for Items 1, 2, 3, or 5, significant findings for those items indicate false positives (i.e., Type I error). Significant findings for Item 4 indicate true positives (i.e., power). The percentages of replications that were detected as showing statistically significant DF are presented in Table A1. As shown in Table A1, power to detect an equal effect size was much lower in conditions of 100 respondents per group than of 1,000 respondents per group. Additionally, even the trivially small DF simulated for Item 4 was often statistically significant with sample sizes of 1,000 per group.

Appendix B

Normal Distribution-Based Indices

Normal Distribution-Based Item DF Effect Size Measures

Additional DF effect size indices of interest can be computed by assuming a standard normal distribution on theta ($M = \bar{\theta}_F$, SD = 1 for the focal group rather than using focal group sample respondents' theta values. It is more efficient and accurate to use a set normal distribution of theta and to take the integral across this distribution, rather than to simulate normally distributed data. In practice, however, the integral across a normal distribution can be accurately approximated by using an appropriate number of (e.g., >10) Gaussian–Hermite quadrature nodes and associated density weights across a discrete normal distribution. The computational formulas are

$$SIDN_i = \sum_{\theta}[ES_{(\theta,\gamma_F)} - ES_{(\theta,\gamma_R)}]w_\theta \tag{11}$$

$$UIDN_i = \sum_{\theta}[|ES_{(\theta,\gamma_F)} - ES_{(\theta,\gamma_R)}|]w_\theta \tag{12}$$

where SIDN_i and UIDN_i are the signed and unsigned item difference in a normal distribution of theta for item *i*. Weights, w_θ, are density weights (i.e., proportions) associated with the level of theta that corresponds to a density function of the normal distribution with a mean equal to the focal group theta mean (cf. Stark et al., 2004). These weights represent the proportion of the respondents expected at each theta level. Multiplying scores by proportions and then summing is equivalent to averaging across the theta distribution (as was done with SIDS and UIDS).

The SIDN_i corresponds directly to the SIDS_i index, and the UIDN_i index corresponds directly to UIDS_i. These indices are presented in the same metric as the SIDS and UIDS, which is that of expected and observed scores. The SIDN can be considered an item-level version of Stark et al.'s (2004) test-level DTFR statistic. The interpretations of the SIDN and UIDN indices are just like those of SIDS and UIDS with the exception of having a hypothetical focal group as the reference. Whereas SIDS indicates the DF encountered by the current sample, the SIDN indicates the DF that would be encountered, on average, by a focal group with normally distributed theta scores. With respect to the taxometric framework, both SIDN and UIDN (a) are item-level indices, (b) are based on an assumed normal distribution, and (c) are presented in the observed score metric. The SIDN allows cancellation across regions of theta, whereas the UIDN does not.

These indices are analogous to polytomous variations of the signed and unsigned area measures (Kim & Cohen, 1991; Kim, Cohen, Alagoz, & Kim, 2007; Raju, 1988, 1990; Wainer, 1993) in the special case in which theta is assumed to be normally distributed. Therefore it is *not* the area between the ES curves per se that is computed in Formulas 6 and 7 but the average difference between the curves across focal group members with a normally distributed theta distribution. The advantage of this approach is that it has a closed-form solution and is not strongly affected by large differences in ESs at extreme levels of theta, in contrast to typical area measures (see Camilli & Shepard, 1994).

Normal Distribution-Based Test DF Effect Size Measures

Additional scale-level indices can be computed as are the sums of SIDN and UIDN. Stark et al. (2004) presented a formula for the integral of the signed difference between ETSs across theta, which is equal to the sum of the j item SIDN indices:

$$Stark\text{'s } DTFR = \sum_{i=1}^{j} SIDN_i \tag{13}$$

A similar index can be computed as the sum of the UIDN indices for the j items,

$$UDTFR = \sum_{i=1}^{j} UIDN_i \tag{14}$$

where UDTFR indicates an unsigned version of the DTFR index. The interpretation of these statistics is much like that of their item-level counterparts. These indices (a) are scale level, (b) use an assumed normal distribution, and (c) are presented in the metric of observed scores. Stark et al.'s (2004) DTFR allows cancellation of DF across both persons and items, whereas the unsigned DTFR (UDTFR) does not allow cancellation across either persons or items.

An additional index takes the absolute value of ETSs for an assumed normal distribution prior to summation across the theta range,

$$UETSDN = \sum_{\theta} [|ETS_{(\theta,\gamma_F)} - ETS_{(\theta,\gamma_R)}|]w_\theta \tag{15}$$

where UETSDN represents the unsigned expected test score difference in a normal distribution. This index (a) is scale-level, (b) uses an assumed normal distribution, (c) allows cancellation across items but not across different regions of theta, and (d) is in the metric of observed scores. UETSDN is a normal distribution-based version of Flowers et al.'s (1999) DTF index, though DTF squares the difference rather than taking the absolute value.

Received June 9, 2009
Revision received December 21, 2009
Accepted January 13, 2010 ■

Correction to Meade (2010)

In the article “A Taxonomy of Measurement Invariance Effect Size Indices” by Adam Meade (*Journal of Applied Psychology, 95* (4), 728-743), there was an error in Formula 6 on page 731 for the pooled standard deviation of the ESSD index. The $SD_{ItemPooled}$ should be:

$$SD_{ItemPooled} = \sqrt{\frac{(N_F - 1)\sigma^2_{ES_{(i|\gamma F)}} + (N_F - 1)\sigma^2_{ES_{(i|\gamma R)}}}{2 * N_F - 2}} \qquad \textbf{(6)}$$

Related to this, in Table 8 on page 739, the ETSSD statistic should have been .094 for the cross cultural comparison and .001 for the Administration Format example.