Published in final edited form as:
J Aging Health. 2012 September ; 24(6): 1044–1076. doi:10.1177/0898264312436877.

Modifying Measures Based on Differential Item Functioning (DIF) Impact Analyses

Jeanne A. Teresi, EdD, PhD[1,2,3], **Mildred Ramirez, PhD**[2,3], **Richard N. Jones, ScD**[4], **Seung Choi, PhD**[5], and **Paul K. Crane, MD, MPH**[6]

[1]Columbia University Stroud Center, New York, NY, USA

[2]Research Division, Hebrew Home at Riverdale, New York, NY, USA

[3]Weill Medical College of Cornell University, New York, NY, USA

[4]Hebrew Rehabilitation Center for Aged, Institute for Aging Research, Boston, MA, USA

[5]Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[6]University of Washington, Seattle, WA, USA

Abstract

Objectives—Measure modification can impact comparability of scores across groups and settings. Changes in items can affect the percent admitting to a symptom.

Methods—Using item response theory (IRT) methods, well-calibrated items can be used interchangeably, and the exact same item does not have to be administered to each respondent, theoretically permitting wider latitude in terms of modification.

Results—Recommendations regarding modifications vary, depending on the use of the measure. In the context of research, adjustments can be made at the analytic level by freeing and fixing parameters based on findings of differential item functioning (DIF). The consequences of DIF for clinical decision making depend on whether or not the patient's performance level approaches the scale decision cutpoint. High-stakes testing may require item removal or separate calibrations to ensure accurate assessment.

Discussion—Guidelines for modification based on DIF analyses and illustrations of the impact of adjustments are presented.

Keywords

differential item functioning; DIF; factorial invariance; impact of DIF; ethnicity; measure modification

Corresponding Author: Jeanne A. Teresi, EdD, PhD, Columbia University Stroud Center and Research Division, Hebrew Home at Riverdale, 5901 Palisade Avenue, Riverdale, New York, 10471, USA Teresimeas@aol.com; JAT61@Columbia.edu.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Introduction

Studies of cognitive processes have shown that several factors affect response to survey items: format, for example, ordering responses horizontally to follow the way an item is read rather than vertically (Mullin, Lohr, Bresnahan, & McNulty, 2000), distracting graphical displays, irrelevant information, and inconsistent presentation of material. Changes in order of item presentation, wording, response options, and recall period may have an impact on item response. For example, pairing items so that context was changed was found to result in higher levels of reported symptoms (Steinberg, 2001). Thus, it has been recommended that such changes be subjected to nonrandomized studies to compare response distributions or to randomized studies to examine measurement properties across versions (U.S. Department of Health and Human Services, 2009).

Modification of measures can have an impact on comparability across settings and groups, for example, altering the percent of respondents admitting to a symptom. However, this may be more of a problem in survey or epidemio-logical research in which single-item indicators are used to determine prevalence. In the case of multiple items that are used to form a severity scale to measure degree, changing an item or two may not have such a great impact on the overall scale. Moreover, the changes are more likely to be problematic in the context of “classical test theory”–developed and applied scales. Using item response theory (IRT) methods and computerized adaptive testing (CAT), well-calibrated items can be used interchangeably. Thus, the exact same item does not have to be administered. Theoretically this permits wider latitude in terms of modification. Items can be equated and mapped to the level of the construct that is being measured. Item responses are combined in a probabilistic way that links the response to a person’s standing on the latent construct. Most measures are not developed or used this way. Rather, items exist independently of the trait and are simply added together. They are related to the trait but cannot be linked directly to a specific point on the measurement scale. Thus, both overall and cut scores can be affected negatively by modifications to measures.

Two general quantitative methods have been used to study measurement equivalence at the item level, with extensions to the scale level. Differential Item Functioning (DIF) analyses using IRT and factorial invariance methods yield similar if not identical parameter estimates (after recalibration or with algebraic manipulations). Both methods examine how items are related to a latent construct and whether this relationship differs across groups that differ in terms of background characteristics such as ethnicity, gender, age, or education. (In this article, the term “trait” or “construct” will refer to a health “state,” “disorder,” or “disability”). For all analyses, the trait (e.g., disability) level is controlled, such that item response is examined across groups at equivalent levels of disability. Potential differences in item performance across groups are examined at each point or interval along the disability continuum because, for example, different ethnic groups may differ in the prevalence of a particular health variable, rendering invalid the comparisons of item parameters that do not take into account differences in the trait distribution. The purpose of this article is to review valid methods and illustrations from the literature of DIF impact related to age, ethnic, and racial groups, with a focus on health-related constructs.

Methods

Qualitative Considerations in Modification of Measures

A best practice for the use of quantitative methods for examination of measures is the iterative commingling of qualitative and quantitative processes (Hambleton, 2006). For example, translations of instruments can be affected by lack of conceptual equivalence in different groups. Qualitative analyses provide information about the reasons for nonequivalence, such as changes in content, format, difficulty of words or sentences, and differences in cultural relevance (Angel, 2006; Johnson, 2006; Manly, 2006). Qualitative methods, used in concert with findings from quantitative methods, are key to eliminating measurement inequality in health disparities research (Krause, 2006). Ideally, cognitive interviews (Jobe & Mingay, 1989; Nápoles-Springer, Santoyo-Olsson, O'Brien, & Stewart, 2006; Willis, 2004) are performed prior to final item construction in order to examine differences in wording that could lead to DIF.

In educational testing, standards and guidelines have long existed for adapting tests, for example, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999); Hambleton and Patsula (1998, 1999); and Van de Vijver and Leung (1997). For example, Malida et al. (2008) outline a taxonomy of five adaptation mechanisms for use with cognitive tests: construct, language, culture, theory, and familiarity/recognition-driven adaptation. Using this formulation, construct refers to situations in which the construct definition may vary across cultures; language refers to lack of semantic equivalent words or idiom-atic expressions. Cultural adaptation refers to experiential relevance of the item to a specific culture. Familiarity adaptation refers to differing levels of exposure to specific tasks in school. In cognitive testing of adults this might refer to tasks such as serial 7's (subtracting 7 from 100 all the way down), which might not have been practiced by those with less education. Theory refers to the inability to identify a translated item at the targeted level of severity or difficulty, for example, the same number of digits. Some literature in educational testing has shown that matched-ability minority test takers as contrasted with majority test takers were less likely to correctly answer easier items (see Scherbaum & Goldstein, 2008). Specifically some groups of Black test takers were disadvantaged, possibly due to cultural differences in the use of "common" words in the easier items.

Hypothesis generation, confirmation and sources of DIF—Because numerous tests of DIF/invariance are performed, in order to avoid false DIF detection, hypotheses generation has been recommended (Roussos & Stout, 1996) together with a review of DIF findings extant with respect to each item. Cognitive interviews are performed prior to item construction in an attempt to avoid wording that might lead to DIF. However, despite such tests, DIF may still appear. Moreover, many legacy items may not have undergone such testing. After DIF is observed, content experts often decide on the source of the DIF and whether or not the DIF results in bias against a particular group. This process has received mixed reviews because content experts are not always proficient in identifying DIF or their sources (see Teresi & Fleishman, 2007). Proposed for use in educational testing, the think aloud protocol (TAP; Ercikan et al., 2010) is a method for examining sources of possible

DIF. Whereas hypotheses are generated in advance of DIF testing in order to guide analyses, the TAPs are used after DIF testing. Translated instruments have been examined to detect differences in word meaning, difficulty, number (length of the words), order, and other potential linguistic differences. Do the thought processes of the respondent correspond to the "hypothesized sources of DIF identified by the expert reviewers"? The questions found most instructive in identifying sources of DIF were, "Do the respondents understand the question?," and "What aspects of the questions were confusing or difficult?" Respondents are asked to think about and explain what is going through their head when they are answering. As an example, the cognitive test item, repetition and recall of "John Brown, 42 Market Street, Chicago," when translated literally in Spanish, was found to be confusing to respondents interviewed in Spanish because numbers in addresses follow the street in the Spanish language, for example, "Market Street, number 42." As another example, an instruction about reporting pain in the last 7 days from the Patient-Reported Outcomes Measurement Information System (PROMIS) items (http://www.nihpromis.org) resulted in confusion on the part of some Puerto Rican older adults with cancer who were interviewed in Spanish; they thought that the items referred to the last 7 days of their life, leading to a recommendation to modify the item stem by inserting the word, "past" into the instruction, "in the last (past) 7 days."

Other approaches to understanding the sources of DIF have recently emerged from the educational testing literature. Some of these methods have limited utility in the context of health-related variables because they examine dichotomous items (Kahraman, de Boeck, & Janssen, 2009), or they examine distractors (multiple choice options on an educational test that are incorrect and are "disproportionately attractive" to some groups of respondents; Penfield, 2010). On the other hand, Penfield (2007) has developed a method for identifying polytomous (ordinal-level) items that "contribute construct-irrelevant variance" to the scale score; this classification scheme may show promise in future efforts to investigate health-related items.

Quantitative Methods Used in Measure Modification

An ultimate goal of the examination of DIF is to either modify the measure by changing or removing items or adjust for DIF. DIF adjustment follows examination of the impact of DIF. Reviewed below are methods that have been used to examine and adjust for DIF impact.

Methods for DIF Adjustment

Analytic-level adjustments—Recommendations for DIF adjustment vary, depending on the use of the measure, for example, for research or clinical purposes. There are more options in the context of research. Measures can be equated if common items are included. By freeing and fixing parameters, based on findings of DIF, DIF-adjusted scales can be used in analyses.

Individual-level adjustments—In the context of CAT, currently one can remove an item with DIF from the bank or flag it as an item that should not be used for certain groups. For example, in the PROMIS CAT platform there is a multiple calibration feature in the software that was designed to handle an item that is shared across projects. A goal is to be

able to use general population calibrations for all items and separate group calibrations for a specific item, for example, “crying” and group, for example, gender. Future developmental work based on Choi (2009) would focus on the capability to account for DIF using group-specific item parameters, including a CAT framework that can account for DIF in real time.

Caveats—. As mentioned above, it is theoretically feasible to provide individually calibrated scores, based on a person’s background characteristics, assuming prior research has been performed and or hypotheses have been generated to provide an understanding of the effect of these variables on response. However, the other side of the coin is that CAT uses subsets of items; if sufficient numbers of items evidence DIF, the initial estimate of the health state (denoted as theta in IRT models) could be compromised. For example, CAT algorithms typically use information (information functions) derived from the item parameters generated from IRT to select initial items for administration. If the information functions vary across different subgroups of people, and that is the basis for selection of initial items, it is possible that certain individuals will not be administered optimal starting items.

Because it may not be feasible or practical to have separate calibrations for all groups and items with DIF, the question arises, what is the impact of DIF? Impact in the context of health-related measures refers to the influence of DIF on the scale score at the aggregate level. Indices of DIF impact include (a) group differences in expected scale scores and changes in group, (b) means, or (c) prevalence rates, before and after DIF adjustment. Individual-level impact measures include (a) changes in the estimates of individual health status scores, and (b) differences between or among versions with respect to classifications in relation to a gold standard. Presented below are methods that have been used to examine DIF impact, followed by examples from the literature. Details of DIF estimation procedures in the context of health measures are reviewed in articles contained in Teresi, Stewart, Morales, and Stahl (2006), and advantages and disadvantages of the methods are summarized in Teresi (2006). The focus below is on methods that are based on latent variable or regression models with invariant parameter estimates.

Measures of DIF Impact

Aggregate measures

Aggregate measures of DIF used in the health literature have been derived from several classes of IRT models. For binary items, the one- and two-parameter logistic models are frequently used (Birnbaum section in Lord & Novick, 1968). The one-parameter logistic model, equivalent to the Rasch (1960/1980) model, and the rating scale model (an extension of the Rasch model for polytomous items) has been used in much DIF testing. The graded response model (Samejima, 1969) is probably the most often used in the literature on DIF applications in health because it permits modeling of polytomous data with multiple ordered response options reflecting symptom severity. The discrimination (a_i) parameter describes how well an item is related to the underlying attribute measured and is equivalent with algebraic manipulation to the factor loading. The severity (location or threshold) parameter (b_i) represents an inflection or cutting point between two adjacent response categories and indicates the difficulty or severity of the item. Methods based on this model include log-

likelihood ratio tests and byproducts of logistic regression with a latent conditioning variable, and methods based on structural equation modeling, including the Multiple Indicators, Multiple Causes (MIMIC) Model (Jöreskog & Goldberger, 1975; Muthén, 1984) and multiple group confirmatory factor analyses (MG-CFA).

An important point introduced earlier is that IRT severity parameters are on the same scale (unit) as the latent health construct, so that a severity or location (*b*) parameter difference of 0.50 represents one half standard deviation on the health trait estimate (θ). Effect sizes at the item level have a relationship to impact measures at the scale level, such that in the context of IRT and ordinal items, large differences between location parameters may result in larger group scale score differences in the middle range of the distribution (Steinberg & Thissen, 2006).

Expected scale scores

Expected item scores and expected total (test or scale) response functions are estimates of magnitude and impact of DIF for two widely used approaches to DIF detection used in health applications: IRT log-likelihood (IRTLR; Thissen, 1991; Thissen, Steinberg, & Wainer, 1993) and the Differential Test Functioning (DTF) index (Raju et al., 2009; Raju, van der Linden, & Fleer, 1995). The method has also been used to accompany ordinal logistic regression with latent variables (Choi, Gibbons, & Crane, 2011) and in the context of MIMIC (Jones, 2006). Group differences in the expected proportion of correct responses (for binary items) or the expected item score (for polytomous-graded or ordinal items) can be considered as magnitude measures. The item-level expected score is the sum (over categories) of the probability of response in Category K, weighted by the category score (e.g., the ordinal code for the category). The total scale expected score can be expressed as the sum over items of the conditional probability of response and is a measure of the impact of DIF on the entire measure (see Collins, Raju, & Edwards, 2000; Kim, Cohen, Alagoz, & Kim, 2007; Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Steinberg & Thissen, 2006; Teresi et al., 2007; Wainer, 1993; Wainer, Sireci, & Thissen, 1991.) The total expected score is called the test response function in the educational testing literature (Lord & Novick, 1968) and relates average expected scale scores to theta (the estimate of health). These scores are typically not weighted by the response frequency; however, such a weight can be applied so that the results reflect the relative frequencies in the sample. Several such impact measures, developed for binary items (Wainer, 1993) and expanded to polytomous items (Kim et al., 2007), have been incorporated into DIF software developed by PROMIS investigators (Choi et al., 2011). These impact measures include the expected impact of DIF on scores in absolute group differences between item true score functions and density-weighted differences between groups. The latter adjusts for the actual distribution of individuals; if few respondents are located at the point where the differences are greatest, the weighted impact will be less.

Differential Measure Functioning: Differential functioning of items and tests (DFIT) Compensatory DIF index

A method for quantifying differences in expected scale scores is DTF (Fleer, 1993; Flowers, Oshima, & Raju, 1999; Oshima, Kushubar, Scott, & Raju, 2009; Oshima, Raju, & Nanda,

2006; Raju et al., 1995); this method has been evaluated (Meade, Lautenschlager, & Johnson, 2007) and applied in health settings (Morales, Flowers, Gutiérrez, Kleinman, & Teresi, 2006; Teresi, Kleinman, & Ocepek-Welikson, 2000; Teresi et al., 2007). Differential functioning at the test level (aggregated DIF impact) is the sum of differential functioning at the item level and indicates how much each item's compensatory (CDIF) contributes to DTF of the whole. DIF in one item can cancel out DIF in another item. Stark et al. (Stark, Chernyshenko, & Drasgow, 2004) provides related impact indices.

Adjusted means—. Several impact measures have been developed in the context of the MIMIC model (a variant of the factor analytic structural equation model, which except for differences in parameterization is equivalent to an IRT model; Muthén & Muthén, 1998-2004), which assumes that all items load on a single underlying latent attribute such as physical function. The MIMIC model can include multiple covariates and studied latent variables. The covariates are modeled as predictors of the latent variable in a multiple regression (structural equation) model. The direct effects of a studied background characteristic (such as ethnic group membership) on the item response (controlling for health status) is the measure of DIF. Jones and Gallo (2002) introduced and Jones (2006) illustrated a MIMIC impact measure comparing the estimated group effects before and after adjustment for DIF. DIF-adjusted and DIF-unadjusted effect estimates can be converted to estimated differences in mean scores on the latent variable.

Individual Impact Measures

Individual impact can be assessed through an examination of changes in health estimates (thetas) with and without adjustment for DIF. The unadjusted thetas are produced from a model with all item parameters set equal for the two groups. The adjusted thetas are produced from a model with parameters that showed DIF based on the IRT results estimated separately (freed) for the groups. The capacity to fix and free parameters based on DIF and compare theta estimates is incorporated into software packages such as IRTPRO (Thissen, 2011). This method permits comparisons of trait (health status) measure estimates that are DIF free to those with parameters estimated without DIF adjustment. This methodology has been used by several authors to examine the individual impact of DIF (e.g., Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Teresi et al., 2009).

Crane and colleagues (Crane, Gibbons, Ocepek-Welikson et al., 2007) extended this general methodology by calculating the difference between naïve scores that ignore DIF and scores that account for all forms of DIF (all of the covariates considered simultaneously) to examine cumulative impact of DIF on individual participants. The distribution of these difference scores is then examined; for individual-level DIF impact, a box-and-whiskers plot of the difference scores is constructed. An obvious query is how much deviance from DIF should be concerning. When a minimally important difference (MID) or minimal clinically important difference (MCID) has been established for the instrument, difference scores are plotted against this value; those larger than that value are termed *relevant individual-level DIF impact*. When such values have not been determined for a scale, some investigators have plotted differences due to DIF against the median standard error of measurement (SEM), and differences larger than that value are termed *salient individual-level DIF impact*.

Individual-level impact may be especially important to clinicians interested in how far a naïve score could be from a score that accounted for DIF for a particular person (see Choi et al., 2011, for an example using the PROMIS anxiety item bank).

The method most often used in relation to DIF impact is the SEM because reference values in terms of the original raw scores (used with other MID methods) are not required. Another method is the use of 0.5 *SD* as an estimate of important differences between DIF-adjusted and DIF-unadjusted health measures. This is equivalent to about 1 SEM, assuming that reliability is at least .75 across the theta continuum (see Norman, Sloan, & Wyrwich, 2003). The flaws of such methods have been reviewed, including the arbitrary nature of the designations (for summaries and recommendations regarding methods for determining minimally or clinically important differences—MID and CMID—see Hays, Brodsky, Johnston, Spritzer, & Hui, 2005; Rejas, Pardo, & Angel-Ruiz, 2008; Revicki, Hays, Cella, & Sloan, 2008).

In the context of IRT-based DIF methods that produce arguably more invariant estimates of the health status measure, methods that produce values that may vary across populations may be less useful (see Revicki et al., 2008). However, even with IRT, distribution-based measures such as the SEM may be problematic. Because the SEM is dependent on an omnibus estimate of reliability (SEM = standard deviation weighted by the square root of the unreliability of the measure), the differing precision of the measure across the latent continuum is not considered. Reise and Haviland (2010) discuss individual change in the context of IRT models. Measures provide different precisions at different positions along the latent continuum; thus, individuals at different trait levels can have different error estimates. For example, individuals positioned at very high and low levels of the trait may not be measured well due to poor information provided. Thus, change measurement will be difficult. On the other hand, theoretically (assuming that model assumptions are met), IRT models can provide interval-level measurement and, thus, better (more invariant) measures of change over time. Members of the PROMIS Statistical Core (Hahn & colleagues; http://www.nihpromis.org) are exploring ways of using individual person standard errors from IRT and CAT in statistical analyses, for example, to calculate change in measures over time to be implemented in the CAT environment (see also Revicki et al., 2008). Using this formulation in the context of impact and defining the unadjusted score as *y1* and the adjusted score as *y2*, the standard error of a difference score would equal $\sqrt{(SE\ y_1)^2 + (SE\ y_2)^2}$. These estimates can be used to conduct a *t* test of the difference between measures for each person. Differences that are not statistically significant would indicate no impact, whereas larger or smaller differences would imply meaningful impact for an individual.

Most of the studies examining DIF impact have been cross-sectional. An area that has received little attention is examining DIF longitudinally. Changes in response over time could occur in the interpretation of response categories or the meaning of the construct could change. For example, severity ratings might be affected by functional decline; as a person ages, a higher threshold of tolerance for symptoms may affect ratings. Most of the longitudinal studies have been in the context of longitudinal invariance and structural equation modeling. This is accomplished by estimating all the loadings across the waves of data in an unrestricted model and comparing the first-order factor loadings when constrained

equal. If the chi-square difference between the two models is significant, this indicates that the loadings are not time invariant. The thresholds as well as loadings can also be constrained as equal. Change in correlated dimensions over time can be modeled in terms of random effects (see Cai, 2010). In the language of IRT, an interaction of time by item difficulty is considered a type of DIF called parameter drift (Bock, Muraki, & Pfeiffenberger, 1988). Item parameters for all items are estimated, constraining each item for each time period to be equal. Then separate runs are performed constraining all parameter estimates equal for both time periods, except the item studied, which is estimated separately for each time period. These runs are performed for each item and the chi-square associated with freeing item parameters examined. If significant, indicating improvement in fit associated with freeing the parameter, longitudinal DIF is observed (Meade, Lautenschlager, & Hecht, 2005). A potential problem with such approaches is that the dependency over time is not modeled, for example, by including a random-effect term to model the repeated measures (Cai). In terms of impact, item characteristic curves and expected scale scores could be compared over time in order to obtain an estimate of impact across waves.

DIF Impact: Examples From the Literature

Reviewed below are examples of DIF impact from the literature. Only studies of DIF that examined impact of health-related constructs in relation to age, race, and/or ethnicity are included. A more systematic review of DIF findings in measures of depression, quality of life, and general health can be found in Teresi, Ramirez, Lai, and Silver (2008).

Aggregate-Level Methods: Changes in Expected Scale Scores and Means

Cognition—The impact of DIF on cognitive measures has been examined in the context of model-based means. For example, Jones and Gallo (2002) examined the Mini-Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975) and found that use of a model that did not account for DIF resulted in a 95% overestimate of the level of cognitive disability for women. In addition, Jones (2006) examined changes (before and after DIF adjustment) in group differences in mean scores on the latent cognitive variable, as measured by the MMSE. In the baseline (non–DIF-adjusted) model, the estimate of the indirect effect (impact) showed a significantly lower level of cognitive impairment for Spanish speakers as contrasted with English speakers; in the DIF-adjusted model, the estimate of the mean difference between groups in the latent cognitive impairment means was less (0 to 0.04 vs. –0.12) and not significant.

Using data from the Spanish and English Neurological Assessment Scales (SENAS; Mungas, Reed, Crane, Haan, & Gonzalez, 2004; Mungas, Reed, Haan, & Gonzalez, 2005; Mungas, Reed, Marshall, & Gonzalez, 2000), four different techniques for detecting DIF were examined (Yang et al., 2011). All methods identified a naming item with a picture of a shrimp as that with a high magnitude of DIF. Differences in expected scale scores were observed, indicating that, on average, the conditional probability of a correct response was higher for Spanish speakers than for English speakers. However, the overall DIF cancelled at the scale level because some items with DIF were easier for English and others for Spanish speakers.

Depression—In a study of the PROMIS depression item bank, several methods were used to assess DIF, including IRTLR, IRT ordinal logistic regression (IRTOLR), MIMIC, and DFIT. Based on review of (a) hypotheses, (b) findings from the literature, and (c) the collective results from the various analyses, two items with high magnitude of DIF were recommended for exclusion from calibration or DIF adjustment: “I felt like crying,” and “I had trouble enjoying things that I used to enjoy.” The item, “I felt I had no energy,” was also flagged as an item that showed DIF across several methods (Teresi et al., 2009). Impact, examined at the aggregate level, was found to be minimal in analyses when mean scale or latent state scores were examined with and without adjustment for DIF. This result was confirmed using the expected scale scores that are based on a separate DIF method. The Differential Test Function (Raju et al., 1995) values (a density-weighted summary of differences between groups in test response functions) were not significant; however, individual impact was larger for some individuals (see below).

As a counterexample, a DIF analyses (Smith et al., 2009) related to the further development of an existing item bank designed for assessing emotional distress in cancer patients did not identify DIF by age or gender, although the use of Rasch analysis might have been a limiting factor in DIF identification. DIF analysis was conducted using a heterogeneous sample of cancer patients (n = 4,919) who had completed a combination of eight psychological distress screening instruments. Items were assessed for DIF by age (two groups based on the median age of 56 years) and gender. The criteria used for the DIF analysis were DIF contrast > 0.50 and $p < .001$ to correct for multiple comparisons. No DIF was observed for any items by gender or age with the exception of HADS-Anxiety (Zigmond & Snaith, 1983) item, “I feel tense or ‘wound-up’,” which was easier to endorse by older patients.

DIF was examined in the 15-item version of the Geriatric Depression Scale (GDS; Yesavage et al., 1982) for age, gender ethnicity, and chronic disease (Broekman et al., 2008). The MIMIC model was used for comparisons of baseline no-DIF and DIF models. Changes in the Modification Index (MI) statistic, a measure of model misfit, were examined and tested as a chi-square statistic. The authors concluded that the bias effects of age, gender, ethnicity, and chronic illness could have an impact on the scale score, with age-related DIF having the greatest impact. However, the authors observed DIF cancellation in that items with DIF in different directions resulted in little aggregate impact at the scale level.

The Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) was examined for DIF using MIMIC to determine the impact of bias through examination of the bias effect on the total score (Grayson, MacKinnon, Jorm, Creasey, & Broe, 2000). In order to identify the trait metric, the reference indicator, “felt depressed,” was constrained to have zero bias. The authors concluded that the impact of DIF on the CES-D was large for many group comparisons, with bias components as high as 64% of the magnitude of the genuine depression effect.

The CES-D (Radloff, 1977) was also examined using an extension of the Mantel–Haenszel method, a proportional odds regression model for polytomous (ordinal) items, with Bonferroni correction (Cole, Kawachi, Maller, & Berkman, 2000). A relatively “DIF-free” 17-item version correlated .99 with the full version, using a cutpoint of > 16 points; the

sensitivity and specificity of the reduced scale against the depression classification varied across cut-points. Item-level bias in favor of Blacks reporting more interpersonal problems (IP) resulted in positive factor-level bias; the proportional odds of Blacks responding higher on the IP subscale were 2.72 times (95% CI: 2.11, 3.51) that of Whites matched on depression. The authors suggest that the two IP items might contribute to the association between "perception of racial prejudice" and depression.

The Mantel-Haenszel method for Ordered Response Categories (Mantel, 1963), an extended standardization procedure using the *z* statistic (Dorans & Schmitt, 1993) was used (Azocar, Areán, Miranda, & Muñoz, 2001) to examine DIF in the Beck Depression Inventory (BDI; Beck, Ward, Mendelsohn, Mock, & Erbaugh, 1961). Latino samples were found to have mean scores up to six points greater than English-speaking samples, given equivalent depression scores. Given scores indicative of low levels of depression, Latino respondents were more likely to endorse the items, "I feel I'm being punished," "I feel like crying," and "I believe I look ugly," resulting in depression scores as much as nine points higher than a nondepressed non-Latino White respondent. Items with DIF artificially increased or, as with "I can't do any work at all," decreased the scale score.

The MIMIC model was used by Gallo et al. (Gallo, Cooper-Patrick, & Lesikar, 1998) to examine DIF impact in the Diagnostic Interview Schedule (DIS; Robins, Helzer, Croughan, & Ratcliff, 1981). Covariates were introduced in order to examine DIF by self-reported race, adjusting for differences in the level of depression and for the effect of MMSE score (Folstein et al., 1975), gender, and marital and educational status. Average scores on the depression measure were higher for older Black respondents in comparison to older Whites. Mean levels of depression were lower for older Blacks than for Whites at the Durham–Piedmont site.

In contrast with some of the findings reviewed above showing DIF in Black and White comparisons, in an examination of DIF associated with depression items in the Primary Care Evaluation of Mental Disorders (PRIME-MD; Spitzer et al., 1994), Hepner and her colleagues (Hepner, Morales, Hays, Edelen, & Miranda, 2008) failed to find any significant DIF among lower-income Black and White women. IRTLRDIF was used to evaluate item performance of the PRIME-MD mood module for Black and White women. Because no DIF impact was observed, it was permissible to proceed with the test of a substantive research question related to whether there were differences between Blacks and Whites in depression levels. A multigroup model was estimated using MULTILOG (Thissen, 1991), in which the item parameters for Black and White women were constrained to be equal. The estimated mean score for Black women was 0.4 *SD* lower than that of White women, providing support for the findings of lower levels of depression among Black women as contrasted with White women.

Function—MIMIC was used by Fleishman, Spector, and Altman (2002) to examine impact by comparing the difference between the coefficients relating age to disability with and without DIF adjustment. Controlling for DIF resulted in at least a 35% change in the effect estimates.

General health—The impact of DIF in the SF-12 (Ware, Kosinski, & Keller, 1996) was examined using the MIMIC (Muthén, 1984) model (Fleishman & Lawrence, 2003). The DIF-adjusted model for physical health reduced education differences, and although the estimated values for race and ethnicity indicators increased, they remained nonsignificant compared with the no-DIF model. The impact was greater for the mental health variable. The effects for Blacks (as contrasted with Whites) was significantly higher in the baseline unadjusted model; however, this effect was reduced by half in the DIF-adjusted model and was not significant. Adjusting for DIF resulted in a significant age effect; as contrasted with the youngest group, those 70 years of age and older were found to have significantly lower emotional well-being estimates.

Using a proportional odds logistic regression method to examine the SF-36 (Ware & Gandek, 1998), an effect-size measure for converting odds ratios to a difference in probabilities between the two groups (Dorans & Holland, 1993) was examined (Perkins, Stump, Monahan, & McHorney, 2006). Scales were rescored excluding items flagged with DIF. The physical function and mental health means for the rescored scales were significantly higher ($p < .001$) than the original scales across comparisons. The rescored general health scale mean was significantly lower than the original for age groups, and several general health scores were significantly different across race groups in the sample of individuals with illness. Removal of items was not recommended because of the relatively short length of the scale.

Quality of life—. The European Organization for Research and Treatment of Cancer Quality of Life (EORTC QLQ-C30; Aaronson et al., 1993) was examined for DIF using the Rasch model (Rasch, 1960/1980). PARSCALE was used for IRT parameter estimation. DIF analysis examined item performance across ethnic groups (Pagano & Gotay, 2005). Impact was assessed by two methods comparing the original QLQ-C30 score with that of a DIF-adjusted score: item removal and partial correlations, using regression analysis. Although 12 (of 30) items demonstrated DIF, none were recommended for removal. Both methods resulted in reduction in differences among ethnic groups in the total score; however, it was the view of the authors that DIF may not have an impact on the measurement properties of the QLQ-C30.

Aggregate Level: Prevalence Estimates

In the U.S.-U.K. cross-national studies, a prevalence study of disability in nursing homes showed that residents in New York nursing homes were more impaired than those in London nursing homes. However, several items were hypothesized to potentially show DIF and were found to have DIF (Teresi, Cross, & Golden, 1989). Removal of such items, for example, stair climbing and bathing, yielded prevalence results that were no longer statistically different between the two cities. Moreover, the new measure was more convergent with a “silver-standard” global rating.

Nonparametric IRT using SIBTEST (Shealy & Stout, 1993) was used to examine the Composite International Diagnostic Interview (CIDI; Kessler, Andrews, Mroczek, Üstün, & Wittchen, 1998). A U.S. general population sample was used to compare Hispanics, Black,

and Whites (Breslau, Javaras, Blacker, Murphy, & Normand, 2008). The percent of lifetime prevalence of depressive episode for comparisons (e.g., Black vs. White; Hispanic vs. White) before and after DIF items were removed were estimated and compared. The lifetime prevalence of a depressive episode, estimated without adjustment was 18.6% for Whites and 12.6% for Blacks ($p = .002$). After DIF adjustment, the rates were somewhat lower for Whites (17.7%) and Blacks (12.4%) but still significantly different ($p = .007$). The prevalence estimates for Whites and Hispanics were similar: 18% before and after item removal.

Exploratory principal components analysis was used to derive the factor structure of the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001) in each of the 4 racial and ethnic groups (non-Hispanic White, African American, Chinese American, and Latino; Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006). A generalized Mantel–Haenszel statistic was used to test for the equality of odds ratios across several strata. A MIMIC model was used to examine covariates (age, sex, and English language ability). DIF impact was assessed using threshold cutoffs derived from the original non–DIF-adjusted scores. The depression prevalence estimates (though significant) were similar, ranging from 15.2% for Chinese Americans to 21.8% for non-Hispanic Whites. Differences of higher magnitude were observed between Chinese American males (11.8%) and females (18.1%). The authors concluded that, based on the findings of similar mean scores and factor structure of the PHQ-9 (Kroenke et al., 2001) in the different groups, it could be used “without adjustment in diverse populations.”

Individual-Level Impact

Although the results related to aggregate impact at the scale level may be mixed, and sometimes trivial due to DIF cancellation (items with DIF were in different directions, thus canceling aggregate impact), impact may be observed at the individual person level, as reviewed below.

Changes in individual scores: Cognition—DIF related to language, self-reported proficiency with written Japanese, age, and educational attainment was assessed in the Cognitive Abilities Screening Instrument (CASI) in two large epidemiologic studies of Japanese American elderly: the *Kame* Project ($n = 1,708$) and the Honolulu-Asia Aging Study (HAAS; $n = 3,148$; Gibbons et al., 2009). IRTOLR was used. Seven CASI items had DIF related to language of testing in *Kame* (registration of one item, recall of one item, similes, judgment, repeating a phrase, reading and performing a command, and following a three-step instruction). DIF impact was determined by subtracting unadjusted IRT scores from IRT scores accounting for DIF. Individual differences related to DIF larger than the median SEM were viewed as salient. In both studies, few items had DIF related to age, and none evidenced DIF related to sex. There were more items with DIF related to educational attainment identified in *Kame* (13 of 27 items) than in HAAS (3 of 35 items). The DIF impact was found to be minimal; however, an earlier study (Crane, van Belle, & Larson, 2004) found considerable DIF in the CASI across ethnic, education, age, and gender groups; DIF adjustment with IRT scores reduced the impact of DIF.

Crane and colleagues (2008) used IRTOLR to examine individual impact associated with DIF in a measure of executive function among 791 Hispanic, White, and African American older adults recruited by the UC Davis Alzheimer's Disease Center. Difwithpar (a hybrid ordinal logistic regression model with a latent conditioning variable) was used to obtain IRT scores accounting for DIF. The impact of DIF for gender, age, education, and ethnicity and language group for individual participants was determined by subtracting their unadjusted IRT score from their IRT score accounting for DIF related to that covariate. IRT scores accounting for DIF related to all demographic variables were also calculated. Differences larger than the median SEM were indicative of meaningful or salient scale-level differential functioning. Accounting for all four sources of DIF simultaneously changed 68 (9%) of individual scores. Most of the impact of DIF was related to education and to race and ethnicity. Some participants had scores that were affected by DIF by as much as one third of a standard deviation. The authors argue that DIF impact of this magnitude is most likely to be problematic when using cutoff scores to determine whether an individual is impaired and could impact the validity of clinical diagnosis. Controlling for ethnicity (language, education, and gender) substantially strengthened relationships with MRI variables; adjusting for age lessened the relationships. Scale-level adjustment for demographic variables like ethnicity and education was beneficial because it removed a variance component from the total ability estimate unrelated to brain structure, such that the brain structure effect remained more salient. Age adjustment had a negative impact on test validity.

Depression—In the analyses of the PROMIS depression item bank, described earlier, individual-level DIF impact was observed using IRTOLR for some people; this result was confirmed using IRTLR (Teresi et al., 2009). An examination of the differences in thetas with and without adjustment for DIF showed that 22.4% of participants changed by at least 0.5 theta (about 0.5 *SD*), of which 5.8% changed by the equivalent of 1 *SD* in the analyses of gender; for education, the figures are 53.5% and 25.1%, and for age 3.8% changed by at least 0.5 *SD*. The impact was in the direction of false positives; using a cutoff of theta ≥ 1, comparable to about 1 *SD* above the mean, 9.5% would be classified as depressed prior to DIF adjustment, but not after adjustment in the analyses of gender; the comparable figures for education and age are 13.6% and 3.4%, respectively. Thus, the individual-level impact was large for at least 100 people.

The BDI (Beck et al., 1961) was examined by Kim et al. (2002) using the two-parameter graded response IRT model. Half of the items on the BDI scale accounted for 80% of the differential test (scale) functioning, and IRT-adjusted cutoff scores reduced considerably the false-negative rate of clinically diagnosed patients with depression who would have been classified as nondepressed without DIF adjustment (Kim et al., 2002). Items accounting for more of the DIF impact were loss of libido, sleep disturbance, weight loss, self-accusation, self-dislike, social withdrawal, somatic preoccupation, irritability, work inhibition, guilt feelings, and sense of failure. An impact on depression classification using cutoff scores adjusted on the basis of the IRT model was observed. The false-negative rate was lowered by more than half (9.6% to 4.6%) using IRT-adjusted cutoff scores, and the percent of those

severely depressed decreased (17.0% to 13.3%); also, the percent of those moderately depressed increased (41.3% to 49.5%).

IRTOLR was used for DIF analyses on the PHQ-9 (Kroenke et al., 2001) when applied to a sample of patients in usual care in two U.S. cities (n = 1,467; Crane et al., 2010). In a sample of HIV-infected people, although the PHQ-9 did not have large amounts of DIF with respect to individual comparisons of African American and White groups, when all sources of DIF were considered, 20% of the mean differences between African Americans and Whites were found to be due to DIF. Ignoring DIF magnified apparent differences between Whites and Blacks (Crane et al., 2010). Six of the nine PHQ-9 items had DIF with respect to at least one covariate. Three items had DIF with respect to race, two with respect to sex, and one with respect to age. Using unadjusted standardized scores, the mean depression score for African Americans was 6.73 points lower than the mean depression score for Whites. These differences were less when using scores that accounted for DIF (5.41 points). There was minimal individual-level DIF impact.

The FACT version 3 (Cella, 1994), Physical Well-Being (PWB), Emotional Well-Being (EWB), Social/Family Well-Being (SFWB), and Functional Well-Being (FWB) scales were examined for DIF using IRTOLR (Crane, Gibbons, Narasimhalu, Lai, & Cella, 2007). A convenience sample of patients with cancer or HIV residing in three cities were studied to examine DIF for language (English vs. Spanish), race (African American vs. White), Hispanic ethnicity (yes/no), education (9 and fewer vs. 10+), literacy (sixth-grade level + vs. lower than sixth grade), and mode of administration (self- vs. interviewer-administration). Differences between each participant's score when accounting for and ignoring DIF were compared to the established MID to determine whether retaining items with DIF would significantly affect scales. Variables that were associated with relevant scale-level differential functioning were PWB: race; EWB: race, ethnicity, language; SFWB: all but gender; and FWB: race, language, education, and mode of administration. The least DIF impact was observed for PWB, and the impact of DIF was observed across all scales for race.

Individual impact: Use of a gold standard criterion—There is little research using gold standard diagnostic measures to assess DIF impact. One example was of a cognitive measure, developed using latent class analysis and designed to exclude items with DIF. The measure was relatively short compared to its original version that had 17 items, which was reduced to 14 or 15 items after removing those with DIF. The measure performed better in terms of receiver operating characteristic (ROC) curves than most screening measures, for example, the MMSE in comparison to diagnostic assessment (Teresi, Kleinman, Ocepek-Welikson, Ramirez et al., 2000; Wilder et al., 1995).

Counter examples—The GDS (Yesavage et al., 1982) was examined using the Rasch models in WINSTEPS (Linacre, 2005). The level of significance of .05 was adjusted for multiple comparisons by Bonferroni correction (Tang, Wong, Chiu, Lum, & Ungvari, 2005). A new version of the scale was developed by removing four items that evidenced DIF. The revised and original versions of the GDS were not significantly different in terms of the area under the ROC curves (AUC).

DIF in the GDS was examined in an older adult home care population (n = 526; Marc, Patrick, Raue, & Bruce, 2008). Of the 15 items evaluated, 12 showed no evidence of either uniform or nonuniform DIF. Three items met the criteria for evidence of bias, with ORs ≥ 2.0 or conversely ≤ 0.50. Only uniform DIF was observed for items by gender, “Are you in good spirits most of the time?,” “Do you feel you have more problems with memory?,” and “Do you feel that your situation is hopeless?” Women were more likely to respond negatively to the first item, and men were more likely to respond positively to the last two items. The item, “Do you feel you have more problems with memory?,” showed uniform DIF by race (non-Whites more likely to respond positively). Two shortened GDS versions were examined for bias in relation to depression diagnosis. The AUC for the 12-item version (0.8681) was not significantly improved over the original 15-item scale (AUC = 0.8716). The 14-item version (deleting “Do you feel you have more problems with memory?”) showed item-level bias for both gender and race. The authors concluded that age, level of education, gender, and race did not have an effect on the measurement properties of the GDS-15 in an older adult home care population.

Discussion

The findings summarized above demonstrate that aggregate scale–level DIF may or may not be observed. If items with DIF are in different directions, overall DIF may cancel, for example, in examination of cognitive tests among Spanish and English speakers (Morales et al., 2006; Orlando-Edelen et al., 2006). Similarly, low impact in terms of expected scale scores was observed in function and emotional distress (Crane, Gibbons, Ocepek-Welikson et al., 2007; Teresi et al., 2007). However, large age differences in adjusted means were observed in a measure of physical function (Fleishman et al., 2002). Some studies of depression measures also found low impact at the aggregate scale level (Broekman et al., 2008; Osborne, Elsworth, Sprangers, Oort, & Hopper, 2004; Pickard, Dalal, & Bushnell, 2006; Zigmond & Snaith, 1983). However, the aggregate impact of DIF in measures of depression was found to be substantial in some studies (Cole et al., 2000; Grayson et al., 2000), including artificially inflated scores for Latinos (in contrast to English speakers; Azocar et al., 2001). Similarly, the greatest impact of DIF was observed in the mental health component of a general health scale (Fleishman & Lawrence, 2003); DIF-adjusted scores rendered the difference between Blacks and Whites nonsignificant. Regardless of aggregate impact, salient DIF may be observed for individuals. In the context of widely used legacy measures, if the cumulative effect of DIF at the scale level does not change estimated ability for individual examinees, then DIF may not present a measurement problem. However, if only a few items have DIF but there is a systematic bias of estimates of health status, this presents a salient measurement problem. In general, for a long scale, removal of an item or two is not going to affect the total score, and usually, not the sensitivity and specificity of a measure. However, with shorter scales as are observed in the measurement of some health constructs, the presence of DIF and adjustments such as item removal could make a difference. Moreover, the cumulative effect of small amounts of DIF could have an impact on individual scores, as discussed above. Although more DIF may be tolerated in longer scales, there may be exceptions, depending on the context and item difficulties. Several studies (Fillenbaum, Heyman, Williams, Prosnitz, & Burchett, 1990; Gurland, Wilder,

Cross, Teresi, & Barrett, 1992; Teresi et al., 1995) found that shorter cognitive measures performed better (evidenced less bias and DIF impact) than longer measures; however, in part this was due to the longer measures containing more difficult items, such that response was affected by educational level. DIF impact may also vary by domain and be more severe with respect to cognitive and depression measures as contrasted with physical function assessments.

From the standpoint of a typical, classical test theory–derived scale, DIF can have a major impact on individuals. For example, at the aggregate level, DIF cancellation may be observed in that DIF goes in different directions for different groups; however, at the individual level impact may be demonstrated. In part this discrepancy can be an artifact of the way impact is examined at the scale level. Expected scale scores may focus on a range of theta that might not reflect the range in which many individuals are positioned. In that case the impact will not be observed. Other methods such as DFIT take the actual observed range of thetas into account; thus, it is important to use the best methods, such as density-weighted indices to assess impact.

Simulation studies offer an approach to characterizing DIF of different kinds and magnitudes and modeling how different inferences (correlations with external variables, prevalence estimates, and individual decision making) are impacted in different DIF conditions. The field is in need of indices of DIF impact that put the magnitude of DIF in the context of the quality of measurement provided by the information. For example, some function of DIF impact over scale information would provide a sense of the relative importance of DIF versus the precision at which a person's underlying trait score can be estimated. Although there are several simulation studies concerning the accuracy of DIF detection methods, none have examined the impact of DIF. The key questions to be addressed by such studies are, "How much DIF makes a difference, and what do we do about it?" Part of the challenge is that there are various purposes of DIF detection with different implied standards for DIF detection. Crane, Gibbons, Ocepek-Welikson et al. (2007) identify three potential users of DIF findings: scale developers, social scientists, and clinicians, each of which would produce different answer to the two key questions. Perhaps most straightforward is the case of a researcher principally interested in group comparisons. In this context, what matters is whether DIF confounds the relationship between an observed risk factor and an observed outcome. Simulation studies could be developed that impose DIF of a known but varying magnitude and type (i.e., uniform, nonuniform), with the impact on exposure–outcome relationships quantified in terms of bias and/or loss of statistical power. A reasonable approach to handling DIF would be to ignore it if it is of a magnitude to produce negligible bias in group comparisons or, as reviewed above, model the effects of DIF using multivariate or multilevel modeling approaches. But standards may differ among different social scientists. Consider a public health researcher, who may be interested in having accurate estimates of prevalence of various health states. Small bias in the central distribution of scale score distributions can produce large differences in prevalence estimates depending on where the decision point or threshold is placed.

Less straightforward is the case of clinicians as well as other high-stakes medical, health, and functional testing situations (e.g., a motor vehicle department making decisions on

suitability of an older adult for keeping an active driver's license). Clinicians wish to know whether decisions based on a test for an individual result are reasonably correct. The concern here is with making individual-level inferences on the basis of a test score. It is probable for many measures used in clinical settings that a larger concern is overall scale precision rather than DIF. But regardless, large DIF may be tolerable if the decision point on the test (i.e., cutpoint) is far removed from the patient's performance level. But as the patient's performance level approaches the decision point, the likelihood of an incorrect decision on the basis of performance increases. What to do about DIF begs the question of how IRT-based scale results should be presented for clinical use. Arguably, a preferred method is one that provides a score for a patient and an interval estimate capturing the range of uncertainty about that score. Simulation studies could demonstrate the potential errors in individual classification against a decision point on a scale in the presence of DIF and serve to validate methods for DIF correction in such contexts.

Conclusions and Recommendations

When high-stakes decision making is involved, it is better to err on the conservative side and consider removal of items with high magnitude of DIF, leading to high impact. As reviewed above, the alternative methods of separate calibrations to adjust for DIF are not mature, and may not be feasible, given the many possible subgroups that might respond differentially to specific items. Based on qualitative review and quantitative analyses, items might be modified to reflect better the original intent and meaning of the item, which may have been altered when applied to different ethnic or educational groups. Efforts such as Toolbox (http://www.nihtoolbox.org), focused on assessment of neurological and behavioral function, and PROMIS (http://www.nihpromis.org), which focuses on the construction of items banks to measure health status, rely on the inclusion of items that have been examined for measurement equivalence. When several methods consistently identify items with a high magnitude of DIF, content experts and investigators are more likely to consider modification. For example, high-magnitude DIF found in two of the PROMIS depression item bank items resulted in their deletion from the item bank, and the shrimp item from a well-know cognitive test was removed as a result of confirmatory DIF analyses (Yang et al., 2011), demonstrating the importance and practical consequences of such analyses.

Acknowledgments

The authors would like to thank the National RCMAR Measurement and Methods Core members for internal reviews of this manuscript, particularly Judy Shea (University of Pennsylvania) and Ron Hays (University of California, Los Angeles). We also acknowledge Steve Wallace (University of California, Los Angeles) for his assistance in assembling this set of papers.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The preparation of this manuscript was supported in part by the following grants: Patient Reported Outcomes Measurement Information System: NCI-U01-AR057971; Centers of Excellence in Health Disparities: NIMHD, P60, MD00206; Claude Pepper Older Americans Independence Center: NIA, P30, AG028741; Resource Centers on Minority Aging Research: NIA P30-AG15272-12S2.

References

Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Takeda F. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute. 1993; 85:365–376. [PubMed: 8433390]

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. APA; Washington, DC: 1999.

Angel RJ. Narrative and the fundamental limitations of quantification in cross-cultural research. Medical Care. 2006; 44(Suppl. 11):31–33. [PubMed: 16365610]

Azocar F, Areán P, Miranda J, Muñoz RF. Differential item functioning in a Spanish translation of the Beck Depression Inventory. Journal of Clinical Psychology. 2001; 57:355–365. [PubMed: 11241365]

Beck AT, Ward CH, Mendelsohn M, Mock J, Erbaugh J. An inventory for measuring depression. Archives of General Psychiatry. 1961; 4:561–571. [PubMed: 13688369]

Bock R, Muraki E, Pfeiffenberger W. Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement. 1988; 25:275–285.

Breslau J, Javaras KN, Blacker D, Murphy JM, Normand SLT. Differential item functioning between ethnic groups in the epidemiological assessment of depression. Journal of Nervous and Mental Disease. 2008; 196:297–306. [PubMed: 18414124]

Broekman BFP, Nyunt SZ, Niti M, Jin AZ, Ko SM, Kumar R, Ng TP. Differential item functioning of the Geriatric Depression Scale in an Asian population. Journal of Affective Disorders. 2008; 108:285–290. [PubMed: 17997490]

Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika. 2010; 75(4):581–612.

Cella, D. Manual for the functional assessment of cancer therapy (FACIT) and functional assessment of HIV infection (FACIT) scales (Version 3). Rush-Presbyterian-St. Luke's Medical Center; Chicago, IL: 1994.

Choi SW. Computerized adaptive testing simulation program for polytomous IRT models. Applied Psychological Measurement. 2009; 33:644–645.

Choi SW, Gibbons LE, Crane PK. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of Statistical Software. 2011; 39(8):1. [PubMed: 21572908]

Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. Journal of Clinical Epidemiology. 2000; 53:285–289. [PubMed: 10760639]

Collins WC, Raju NS, Edwards JE. Assessing differential item functioning in a satisfaction scale. Journal of Applied Psychology. 2000; 85:451–461. [PubMed: 10900818]

Crane PK, Gibbons LE, Narasimhalu K, Lai JS, Cella D. Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. Quality of Life Research. 2007; 16:101–114. [PubMed: 17111233]

Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, Teresi JA. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Quality of Life Research. 2007; 16:69–84. [PubMed: 17554640]

Crane PK, Gibbons LE, Willig JH, Mugavero MJ, Lawrence ST, Schumacher JE, Crane HM. Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). AIDS Care. 2010; 22:874–885. [PubMed: 20635252]

Crane PK, Narasimhalu K, Gibbons LE, Pedraza O, Mehta KM, Tang Y, Mungas DM. Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. Journal of the International Neuropsychological Society. 2008; 14:746–759. [PubMed: 18764970]

Crane PK, Van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004; 23:241–256. [PubMed: 14716726]

Dorans, NJ.; Holland, PW. DIF detection and description: Mantel Haenszel and standardization.. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Lawrence Erlbaum; Hillsdale, NJ: 1993. p. 35-66.

Dorans, NJ.; Schmitt, AP. Constructed response and differential item functioning: A pragmatic approach.. In: Bennett, RE.; Ward, WC., editors. Construction versus choice in cognitive measurement. Lawrence Erlbaum; Hillsdale, NJ: 1993. p. 135-165.

Ercikan K, Arim R, Law D, Domene J, Gagnon F, Lacroix S. Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. Educational Measurement: Issues and Practice. 2010; 29:24–35.

Fillenbaum GG, Heyman A, Williams K, Prosnitz B, Burchett B. Sensitivity and specificity of standardized screens of cognitive impairment and dementia among elderly black and white community residents. Journal of Clinical Epidemiology. 1990; 43:650–660.

Fleer PF. A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. Dissertation Abstracts International. 1993; 54(04B):2266.

Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: True differences or differential item functioning? Medical Care. 2003; 41(Suppl. 3):75–86.

Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. Journals of Gerontology: Social Sciences. 2002; 57:275–284.

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999; 23:309–332.

Folstein M, Folstein S, McHugh P. Mini-Mental State: A practical guide for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]

Gallo JJ, Cooper-Patrick L, Lesikar S. Depressive symptoms of Whites and African Americans aged 60 years and older. Journals of Gerontology. 1998; 53:277–285.

Gibbons LE, McCurry S, Rhoads K, Masaki K, White L, Borenstein AR, Crane PK. Japanese–English language equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans. International Psychogeriatrics. 2009; 21:129–137. [PubMed: 18947456]

Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item bias in the Center for Epidemiological Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. Journal of Gerontology: Psychological Sciences. 2000; 55:273–282.

Gurland BJ, Wilder DE, Cross PE, Teresi JA, Barrett VW. Screening scales for dementia: Toward reconciliation of conflicting cross-cultural findings. International Journal of Geriatric Psychiatry. 1992; 7:105–113.

Hambleton RK. Good practices for identifying differential item functioning. Medical Care. 2006; 44(Suppl. 11):182–188. [PubMed: 16434918]

Hambleton RK, Patsula L. Adapting tests for use in multiple languages and cultures. Social Indicators Research. 1998; 45:153–171.

Hambleton RK, Patsula L. Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. Journal of Applied Testing Technology. 1999; 1:1–30.

Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui K-K. Evaluating the statistical significance of health related quality-of-life change in individual patients. Evaluation and the Health Professionals. 2005; 28:160–171.

Hepner KA, Morales LS, Hays RD, Edelen MO, Miranda J. Evaluating differential item functioning of the PRIME-MD mood module among impoverished Black and White women in primary care. Women's Health Issues. 2008; 18:53–61. [PubMed: 18069001]

Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. Journal of General Internal Medicine. 2006; 21:547–552. [PubMed: 16808734]

Jobe JB, Mingay DJ. Cognitive research improves questionnaires. American Journal of Public Health. 1989; 79:1053–1055. [PubMed: 2751028]

Johnson T. Methods and frameworks for cross-cultural measurement. Medical Care. 2006; 44(Suppl. 11):17–20.

Jones RN. Identification of measurement differences between English and Spanish language versions for the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care. 2006; 44(11, Suppl. 3):124–133. [PubMed: 16434911]

Jones RN, Gallo JJ. Education and sex differences in the Mini-Mental Status Examination: Effects of differential item functioning. Journal of Gerontology: Psychological Sciences. 2002; 57:548–558.

Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association. 1975; 10:631–639.

Kahraman N, De Boeck P, Janssen R. Modeling DIF in complex response data using test design strategies. International Journal of Testing. 2009; 9:151–166.

Kessler RC, Andrews G, Mroczek D, Üstün TB, Wittchen H-U. The World Health Organization Composite International Diagnostic Interview Short Form (CIDI-SF). International Journal of Methods in Psychiatric Research. 1998; 7:171–185.

Kim S, Cohen AS, Alagoz C, Kim S. DIF detection and effect size measures for polytomously scored items. Journal of Educational Measurement. 2007; 44:93–116.

Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. Psychology and Aging. 2002; 17(3):379–391. [PubMed: 12243380]

Krause N. The use of qualitative methods to improve quantitative measures of health-related constructs. Medical Care. 2006; 44(11, Suppl. 3):34–38.

Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine. 2001; 169:606–613. [PubMed: 11556941]

Linacre, JM. Mesa Press; Chicago, IL: 2005. Winsteps: Rasch model computer programs..

Lord, FM.; Novick, MR.; Birnbaum, A. Statistical theories of mental test scores. Addison-Wesley; Reading, MA: 1968.

Malida M, Van de Vijver FJR, Srinivasan K, Tranlser C, Sukumar P, Rao K. Adapting a cognitive test for different culture: An illustration of qualitative procedures. Psychology Science Quarterly. 2008; 50:453–468.

Manly JJ. Deconstructing race and ethnicity: Implications for measurement of health outcomes. Medical Care. 2006; 44(Suppl. 11):10–16.

Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. Journal of the American Statistical Association. 1963; 58:690–700.

Marc LG, Patrick MS, Raue J, Bruce ML. Screening performance of the 15-Item Geriatric Depression Scale in a diverse elderly home care population. American Journal of Geriatric Psychiatry. 2008; 16:914–921. [PubMed: 18978252]

Meade A, Lautenschlager G, Hecht JE. Establishing measurement equivalence and invariance in longitudinal data with item response theory. International Journal of Testing. 2005; 5:279–300.

Meade A, Lautenschlager G, Johnson E. 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. 2007; 31:430–455.

Morales LS, Flowers C, Gutiérrez P, Kleinman M, Teresi JA. Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) framework. Medical Care. 2006; 44(11, Suppl. 3):143–151.

Mullin PA, Lohr KN, Bresnahan BW, McNulty P. Applying cognitive design principles to formatting HRQOL instruments. Quality of Life Research. 2000; 9(1):13–27. [PubMed: 10981203]

Mungas D, Reed BR, Crane PK, Haan MN, Gonzalez H. Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. Psychological Assessment. 2004; 16(4):347–359. [PubMed: 15584794]

Mungas D, Reed BR, Haan MN, Gonzalez H. Spanish and English neuropsychological assessment scales: Relationship to demographics, language, cognition, and independent function. Neuropsychology. 2005; 19(4):466–475. [PubMed: 16060821]

Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology. 2000; 14(2): 209–223. [PubMed: 10791861]

Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984; 49:115–132.

Muthén, LK.; Muthén, BO. Mplus user's guide. 3rd ed.. Authors; Los Angeles, CA: 1998-2004.

Nápoles-Springer AM, Santoyo-Olsson J, O'Brien H, Stewart AL. Using cognitive interviews to develop surveys in diverse populations. Medical Care. 2006; 44(Suppl. 3):21–30.

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. Medical Care. 2003; 41:582–592. [PubMed: 12719681]

Orlando-Edelen M, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. Medical Care. 2006; 44(11, Suppl. 3):134–142.

Osborne RH, Elsworth GR, Sprangers MAG, Oort FJ, Hopper JL. The value of the Hospital Anxiety and Depression Scale (HADS) for comparing women with early onset breast cancer with population-based reference women. Quality of Life Research. 2004; 13:191–206. [PubMed: 15058800]

Oshima, TC.; Kushubar, S.; Scott, JC.; Raju, NS. DFIT8 for Window user's manual: Differential functioning of items and tests. Assessment Systems Corporation; St. Paul, MN: 2009.

Oshima TC, Raju NS, Nanda AO. A new method for assessing the statistical significance of the differential functioning of items and tests (DFIT) framework. Journal of Educational Measurement. 2006; 43:1–17.

Pagano IS, Gotay CC. Ethnic differential item functioning in the assessment of quality of life in cancer patients. Health and Quality of Life Outcomes. 2005; 3:60–69. [PubMed: 16209720]

Penfield RD. An approach for categorizing DIF in polytomous items. Applied Measurement in Education. 2007; 20:335–355.

Penfield RD. Distinguishing between net and global DIF in polytomous items. Journal of Educational Measurement. 2010; 47:129–149.

Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 Health Survey. Quality of Life Research. 2006; 15:331–348. [PubMed: 16547771]

Pickard AS, Dalal MR, Bushnell DM. A comparison of depressive symptoms in stroke and primary care: Applying Rasch models to evaluate the Center for Epidemiologic Studies-Depression Scale. Value in Health. 2006; 9:59–64. [PubMed: 16441526]

Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401.

Raju NS, Fortmann-Johnson KA, Kim W, Morris SB, Nering ML, Oshima TC. The item parameter replication method for detecting differential functioning in the DFIT framework. Applied Measurement in Education. 2009; 33:133–147.

Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995; 19:353–368.

Rasch, G. Probabilistic models for some intelligence and attainment tests. University of Chicago Press; Chicago, IL: 1980. (Original work published 1960)

Reise SP, Haviland MG. Item response theory and the measurement of clinical change. Journal of Personality Research. 2010; 84:228–238.

Rejas J, Pardo A, Angel-Ruiz M. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. Journal of Clinical Epidemiology. 2008; 61:360–356.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008; 61:102–109. [PubMed: 18177782]

Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. Archives of General Psychiatry. 1981; 38:381–389. [PubMed: 6260053]

Roussos LA, Stout WF. Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement. 1996; 33:215–230.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 34:100–114.

Scherbaum CA, Goldstein HW. Examining the relationship between race-based differential item functioning and item difficulty. Educational and Psychological Measurement. 2008; 68:537–553.

Shealy RT, Stout WF. 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. 1993; 58:159–194.

Smith AB, Rush R, Wright P, Stark D, Velikova G, Sharpe M. Validation of an item bank for detecting and assessing psychological distress in cancer patients. Psycho-Oncology. 2009; 18:195–199. [PubMed: 18677714]

Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV III, Hahn SR, Johnson JG. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. Journal of American Medical Association. 1994; 272(22):1749–1756.

Stark S, Chernyshenko OS, Drasgow F. 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. 2004; 89:497–508. [PubMed: 15161408]

Steinberg L. The consequences of pairing questions: Context effects in personality measurement. Journal of Personality and Social Psychology. 2001; 81:332–342. [PubMed: 11519936]

Steinberg L, Thissen D. Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods. 2006; 11:402–415. [PubMed: 17154754]

Tang WK, Wong E, Chiu HFK, Lum CM, Ungvari GS. The Geriatric Depression Scale should be shortened: Results of Rasch analysis. International Journal of Geriatric Psychiatry. 2005; 20:783–789. [PubMed: 16035120]

Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Medical Care. 2006; 44(Suppl. 11):152–170.

Teresi J, Cross P, Golden R. Some applications of latent trait analysis to the measurement of ADL. Journal of Gerontology: Social Sciences. 1989; 44(5):196–204.

Teresi JA, Fleishman JA. Differential item functioning and health assessment. Quality of Life Research. 2007; 16(Suppl. 1):33–42. [PubMed: 17443420]

Teresi JA, Golden RR, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. Journal of Clinical Epidemiology. 1995; 48:473–483. [PubMed: 7722601]

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. Statistics in Medicine. 2000; 19:1651–1683. [PubMed: 10844726]

Teresi JA, Kleinman M, Ocepek-Welikson K, Ramirez M, Gurland B, Lantigua R, Holmes D. Applications of item response theory to the examination of the psychometric properties and differential item functioning of the CARE Dementia Diagnostic Scale among samples of Latino, African-American and White non-Latino Elderly. Research on Aging. 2000; 22:738–773.

Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Cella D. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measure of physical functioning ability and general distress. Quality Life Research. 2007; 16:43–68.

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke J, Crane PK, Jones RN, Cella D. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51(2):148–180. [PubMed: 20336180]

Teresi JA, Ramirez M, Lai J-S, Silver S. Occurrences and sources of differential item functioning (DIF) in patient reported outcomes measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychology Science Quarterly. 2008; 50:538–612. [PubMed: 20165561]

Teresi JA, Stewart AL, Morales L, Stahl S. Measurement in a multi-ethnic society: Overview to the special issue. Medical Care. 2006; 44(Suppl. 11):3–4.

Thissen, D. MULTILOGTM user's guide. Multiple, categorical item analysis and test scoring using item response theory. Scientific Software, Inc.; Chicago, IL: 1991.

Thissen, D. IRTPRO: Beta features and Operations. Scientific Software, Inc.; Chicago, IL: 2011.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models.. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Lawrence Erlbaum; Hillsdale, NJ: 1993. p. 67-113.

U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for industry patient reported outcome measures: Use in medical product development to support labeling claims. Author; Washington, DC: 2009.

Van de Vijver, FJR.; Leung, K. Methods and data analysis for cross-cultural research. Sage; Thousand Oaks, CA: 1997.

Wainer, H. Model-based standardization measurement of an item's differential impact.. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Lawrence Erlbaum; Hillsdale, NJ: 1993. p. 123-135.

Wainer H, Sireci SG, Thissen D. Differential testlet functioning: Definitions and detection. Journal of educational measurement. 1991; 28(3):197–219.

Ware JE, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. J Clin Epidemiol. 1998; 51(11):903–12. [PubMed: 9817107]

Ware JE, Kosinski M, Keller SD. A 12-item, short-form health survey: Construction of scales and preliminary tests of reliability and validity. Medical Care. 1996; 34:220–233. [PubMed: 8628042]

Wilder D, Cross P, Chen J, Gurland B, Lantigua R, Teresi J, Encarnacion P. Operating characteristics of brief screens for dementia in a multicultural population. The American Journal of Geriatric Psychiatry. 1995; 3:1–12.

Willis, GB. Cognitive interviewing revisited: A useful technique, in theory?. In: Presser, S.; Rothgeb, JM.; Couper, MP., editors. Methods for testing and evaluating survey questionnaires. John Wiley; New York, NY: 2004. p. 23-43.

Yang FM, Heslin KC, Mehta KM, Yang C-W, Jones RN, Ocepek-Welikson K, Teresi JA. A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. Psychological Test and Assessment Modeling. 2011; 53(4):440–460. [PubMed: 23471423]

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. Development and validation of a geriatric depression screening scale: A preliminary report. Journal of Psychiatric Research. 1982; 17:37–49. [PubMed: 7183759]

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandanavica. 1983; 67:361–370.

Recommendations

Based on the above review, the following recommendations are given.

1. Items that have consistently demonstrated high magnitude of DIF that impacts the scale score should be removed. This practice should be rare in professionally developed health assessment measures.
2. Items that show DIF of high magnitude for some groups and are deemed important to the conceptual map or clinically relevant can be retained, but modified versions of the item can be included either as additional test items or as substitutes.
3. Slight modifications (adding an additional clarifying word or phrase) are acceptable. For example “vigorous activities” is an item that shows DIF. A modification is to define in parentheses the meaning of “vigorous.”
4. Modifications of a few items within a longer scale, for example, 2 to 3 items in a 20-item scale is less likely to have a great impact on a scale.
5. In the context of shorter scales with a mixture of more or less severe items (heterogeneous scales), modifications may have more of an impact.
6. In the context of item banks, ideally separate group calibrations should be used when possible, based on findings of DIF. However, with small subsets of items administered, it may be best to be conservative and remove from the calibrated bank items with a high magnitude of DIF that are likely to impact the estimate of health status.

In summary, although research studies can adjust for DIF at the analytic level, current methods do not permit such adjustments at the individual level such as in the clinical use of a measure. Both theory and evidence from the literature indicates that items that are not equivalent will result in inability to compare means across groups, and in inaccurate classification at the individual level. Although modifications of measures can have consequences in the ways reviewed above, in the view of the authors, it is not reasonable to adhere to a rigid rule of “no modifications.” Modifications of items are sometimes necessary, and thus, recommended.