Psychology Science Quarterly, Volume 50, 2008 (4), pp. 538-612

Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health

JEANNE A. TERESI[1,2], MILDRED RAMIREZ[2], JIN-SHEI LAI[3] & STEPHANIE SILVER[2]

Abstract

Examination of the equivalence of measures involves several levels, including conceptual equivalence of meaning, as well as quantitative tests of differential item functioning (DIF). The purpose of this review is to examine DIF in patient-reported outcomes. Reviewed were measures of self-reported depression, quality of life (QoL) and general health. Most measures of depression contained large amounts of DIF, and the impact of DIF at the scale level was typically sizeable. The studies of QoL and health measures identified a moderate amount of DIF; however, many of these studies examined only one type of DIF (uniform). Relative to DIF analyses of depression measures, less analysis of the impact of DIF on QoL and health measures was performed, and the authors of these analyses generally did not recommend remedial action, with one notable exception. While these studies represent good beginning efforts to examine measurement equivalence in patient-reported outcome measures, more cross-validation work is required using other (often larger) samples of different ethnic and language groups, as well as other methods that permit more extensive analyses of the type of DIF, together with magnitude and impact.

Key words: Differential Item Functioning (DIF), measurement equivalence, patient-reported outcomes, quality of life, depression, general health

[1] Columbia University Stroud Center, Faculty of Medicine and New York State Psychiatric Institute

[2] Research Division, HHAR, 5901 Palisade Avenue, Riverdale, New York 10471, USA, Tel.: 718-581-1139, Fax: 718-543-2477, email: Teresimeas@aol.com, jat61@columbia.edu

[3] Center on Outcomes, Research and Education (CORE); Evanston Northwestern Healthcare, and Northwestern University

Introduction

A series of cross-national studies of mental health conditions among community residents, conducted in the 1960s and 1970s (e.g., Gurland, Fleiss, Cooper, Kendell & Simon, 1969) identified differences among countries in terms of prevalence rates. A question arose regarding the reasons for the differences, and whether measurement bias could be partially responsible. One result of these studies was the development of standardized methods for assessment of physical and mental health states (e.g., Copeland et al., 1976; Golden, Teresi, & Gurland, 1984; Gurland et al., 1972; Spitzer, Fleiss, Burdock, & Hardesty, 1964). More recently, these efforts at standardization have resulted in initial attempts (e.g., Bode et al., 2006; Fliege et al., 2005; Hahn, Cella, Bode, Gershon, & Lai, 2006; Lai, Cella, Chang, Bode, & Heinemann, 2003; Lai et al., 2005; Reeve et al., 2007; Ware et al., 2003) to develop item banks that can be used in computerized adaptive testing (CAT). Because relatively smaller subsets of items are used to establish health status in CAT, it is necessary to ensure that the item bank is acceptable from the perspective of measurement equivalence.

At the same time, there has been growing concern about disparities in access to and delivery of health services, possibly resulting in differential outcomes associated with care (Bloche, 2004; National Research Council, 2004; Smedley, Stith & Neslon (Eds.), 2003; Steinbrook, 2004). Health care decisions are often made based on assessments of the health status of individuals; however, as illustrated in an edited volume of reviews (Skinner, Teresi, Holmes, Stahl & Stewart, 2001) evidence of the cultural equivalence of health-related measures is sparse. The major goal of a recently published special issue of *Medical Care* (Teresi, Stewart, Morales & Stahl, 2006) was to provide state-of-the-art overviews of both qualitative and quantitative methods that can be used to examine measurement equivalence. A major quantitative method for the examination of cultural equivalence is differential item functioning (DIF). Because of the increasing diversity observed in many societies, such analyses are becoming central to measurement development and evaluation. The purpose of this review article is to summarize the findings with respect to DIF in measures of self-reported depression, quality of life and general health. Detailed reviews are presented in an accompanying table.

Although examination of the equivalence of measures involves several levels, including conceptual equivalence of meaning, this review will focus narrowly on methods and results based on analysis of DIF. In addition to focusing only on quantitative methods, excluded from formal review in the accompanying table are analyses of factorial invariance, which is another method for examining measurement equivalence. A discussion of the similarities and differences of these two approaches (factorial invariance and DIF analyses) is beyond the scope of this review, but is summarized in several articles (McDonald, 2000; Meade & Lautenschlager, 2004; Mellenbergh, 1994; Millsap & Everson, 1993; Raju, Laffitte & Byrne, 2002; Reise, Widaman & Pugh, 1993; Takane & De Leeuw, 1987; Teresi, 2006a). The focus of this article will be on issues critical to the examination of DIF, as well as a review of DIF analyses in patient reported outcomes in selected areas. Definitions of the concepts used within this article will be provided as a means of orienting the reader to the topic.

Definition of DIF

DIF involves the evaluation of conditional relationships between item response and group membership. Groups should be selected for study based on theoretical considerations that include whether or not the construct studied is hypothesized to have the same conceptual meaning across groups. For example, if the construct studied is a specific type of pain, clinical experts should decide if this is best measured by a disease-specific or generic scale. If disease-specific, it makes little sense to study that scale for DIF with respect to different disease groups because the construct itself was intended to be different across groups. On the other hand, if a theoretical argument can be advanced that a scale, e.g., a health-related quality of life (HRQoL) subscale should measure the same unidimensional construct across groups that differ in education, literacy or type of disease, then the scale should be studied to insure that DIF is of low magnitude.

As an illustration of a definition of DIF: a randomly-selected person of low literacy with low perceived HRQoL should have the same chance of responding in the low HRQoL direction to an item measuring HRQoL as would a randomly selected individual also with low HRQoL, but who is of high literacy. For this example, uniform DIF indicates that the DIF is in the same direction across the HRQoL continuum, while non-uniform DIF means that the direction of DIF is different, depending on the level of HRQoL.

Magnitude of DIF refers to the degree of DIF, and can be measured by examining parameters or statistics associated with the method, for example, the odds ratio, beta coefficient or increment in R-square associated with the DIF term for the studied item. An important point is that in mental and physical health assessment, the pool of items is limited so that items cannot be discarded as easily as in educational testing. Because DIF detection methodologies are influenced by sample size, many items may show significant DIF for at least one comparison, even after adjustment for multiple group comparisons; however, such statistically significant DIF may not be clinically meaningful. It is thus critical to assess the magnitude of DIF in order to assess DIF saliency. Internal impact goes beyond the item level to determine the impact of DIF on the entire measure or scale. Impact can be assessed at the aggregate level by examining the relationship between the expected scale score and the disability or quality of life estimate; for example, how much do mean group differences in total score distributions change with and without inclusion of the items with DIF? Another example is the impact of DIF on the relationships of demographic characteristics with health variables (Crane, Gibbons, Jolley & Van Belle, 2006; Fleishman & Lawrence, 2003; Fleishman, Spector & Altman, 2002; Morales, Flowers, Gutierrez, Kleinman & Teresi, 2006; Teresi, Cross & Golden, 1989). DIF may also influence the relationship between patient-reported health variables and predicted outcomes such as access to care, functional decline and morbidity. This latter relationship has been referred to as external impact, predictive validity or predictive scale bias, and may be examined in terms of predictive values and regression coefficients. The impact measures just described are all at the aggregate or group rather than individual level. The impact on specific individuals can also be examined.

One method for DIF adjustment is the removal of items that contribute to overall DIF; this is not necessarily the best procedure because most analyses have been of pre-existing relatively short scales. Item removal can thus result in change of meaning of the construct, imbalance of severe and less severe indicators and lowered reliability (see Teresi, 2006b; Hambleton, 2006). Moreover, some items may show DIF cancellation, and their removal could bias a test because some favor one group and others the other group. Borsboom, Mellenbergh and Van Heerden (2002) and Borsboom (2006) discuss the idea that within-group comparison leads to absolute rather than relative bias; items showing absolute (but not relative) bias may not need to be removed. On the other hand, in large-scale item banking projects, in which the goal is to construct relatively DIF-free item sets, removal of items with a large magnitude of consistently identified DIF may be warranted. When selection and treatment decisions are based on individual person-assessments, the presence of DIF in the measure can result in bias and negative impact of DIF. These decisions must be balanced in the context of a conceptual map drawn by content experts who help to determine the relative salience of each item for the intended construct.

Different methods of DIF detection

Presented briefly are several methods for DIF detection that were used in the articles reviewed below. Methods can be categorized broadly as parametric or non-parametric, and differ in terms of whether the conditioning patient-reported outcome variable is based on a latent variable or observed score.

Non-parametric methods

Non-parametric contingency table approaches include the Mantel-Haenszel chi-square method (M-H; see Holland & Thayer, 1988; Dorans & Holland, 1993). The M-H method examines whether the odds of a symptomatic response within each score group on the measure is the same across groups. This method was used by two authors reviewed in Table 1 (Azocar, Areán, Miranda & Muñoz, 2001; Cole, Kawachi, Maller & Berkman, 2000) to examine depression measures. An extension of this method using a variant of the gamma statistic (Goodman & Kruskal, 1954) for polytomous responses was used by Bjorner and colleagues (1998) to examine the SF-36 Health Survey (Ware, Gandek & The IQOLA Project Group, 1994), and by Groennvold and colleagues (1995). Using the M-H method, a common odds ratio (which tests whether or not the likelihood of item symptom response is the same across disability groups) also can be used to construct a DIF magnitude measure. Odds are converted to log odds and various transformations provide interpretable magnitude measures.

An advantage of such methods is that few assumptions are required. However, similar to the Rasch model (Rasch, 1980; described below), this method may not be optimal if the discrimination parameters vary across groups (see Bock, 1993). Moreover, the M-H method uses an observed score treated as a categorical variable rather than the theoretically preferred latent conditioning variable.

Parametric methods using observed scores

Related to the M-H method is the parametric contingency-table approach to the examination of DIF, based on logistic regression (LR; Swaminathan & Rogers, 1990). LR examines whether the odds of admitting to a symptom are different between two groups. Item response is predicted from total observed scores, group status and the interaction of group by the total score. Like M-H, this method has traditionally used a summary conditional disability measure based on observed scores; however, Crane and colleagues (2004, 2006), have developed a method that incorporates a latent conditioning variable based on IRT estimation. LR has been expanded to include ordinal logistic regression (OLR) to accommodate polytomous data. A likelihood test, distributed as Chi-square with separate tests for uniform and non-uniform DIF has been recommended (Jodoin & Gierl, 2001). Purification has also been recommended (Camilli & Shepard, 1994), in which a corrected or unbiased estimate of DIF is achieved by removing items with DIF from the total score (but retaining the studied item) before the final DIF analysis.

LR procedures yield estimates of odds ratios that provide information about the direction and magnitude of the DIF. For example, inclusion of the group variable (for a test of uniform DIF) permits computation of the exponent of the regression coefficient associated with group membership. In addition, at each step in the model building, a corresponding estimate of effect size value is an R^2 difference from the OLR models that can be applied to both binary and ordinal items (Gelin & Zumbo, 2003; Zumbo, 1999). The beta coefficient also can be evaluated for significance and for magnitude.

Advantages of LR include ability to model multidimensional data and capability to include covariates. A disadvantage is the use of the observed conditioning variable. While it is possible to use a latent conditioning variable (see Crane, Van Belle & Larson, 2004), this is infrequently applied in practice, and adds to the complexity of the analyses (see Millsap, 2006). Crane and colleagues (Crane, Gibbons, Narasimhalu, et al., 2007; Crane, Gibbons, Ocepek-Welikson, et al., 2007) used the latent conditioning OLR method in the analyses of a quality of life and a general health measure reviewed in Table 1. LR with an observed conditioning variable was used to cross-validate results in the paper by Petersen and colleagues (2003) to examine a measure of quality of life. It was also used in the studies by Cole and colleagues (2000) to examine depression items, and by Scott and colleagues (2006a, 2007) and Perkins and colleagues (2006), examining quality of life and health, respectively.

Latent variable parametric approaches

Estimates of disability in latent variable models are not based on observed scores. Latent variable methods include MIMIC (Muthén, 1984) and various item response theory (IRT) models such as the Rasch, one parameter IRT model or the IRTLR tests based on the two or three-parameter IRT models. Most of these models can be linked to IRT; thus a brief explication of IRT follows. According to the IRT model, an item shows DIF if people from different subgroups but at the same level on the underlying construct measured, have unequal probabilities of responding symptomatically to a particular item.

Item scores are related to the level of the underlying construct, e.g., depression, by functions that provide an estimate of the probability of occurrence of each possible score on an item for a randomly selected individual of given disorder, disability or symptomatology level. In most applications in psychology and physical health, one or two parameters are estimated: the item difficulty (b_i) is the point on the total symptomatology continuum where the probability of a specific symptom response is .5. In applications in which symptoms are scored in a disordered direction, a high b means that the item maximally discriminates (separates symptomatology levels or groups) at a higher or more severe level of symptomatology. High *b's* are characteristic of items that are positively responded to by individuals with more symptomatology, so that relative to items with lower *b's*, individuals have to be at greater levels of symptomatology before they will have a 50% chance of endorsing the item. The difficulty (severity) parameter is tested for uniform DIF. The two-parameter logistic (for binary items; Lord and Novick, 1968) and graded response (Samejima, 1969) or generalized partial credit models (Muraki, 1992) (for polytomous items) estimate a discrimination parameter that is used in tests of non-uniform DIF.

The one parameter (1-PL) Rasch (Rasch, 1960) model (used in many of the DIF studies reviewed here) does not incorporate a discrimination parameter, and can be used to examine uniform DIF when the non-uniform DIF is not of concern. Rasch models are among the most popular methods for use in development and evaluation of health measures; as a byproduct, DIF analyses can be performed. The basic concept of this approach is to compare the item locations between two groups (i.e., reference versus focal or studied groups) using a t-test. Assume that item i has two difficulty estimates (d_{i1} and d_{i2}) for groups 1 and 2 with associated error, S_{i1} and S_{i2}, respectively. The formula to test for DIF will be $t_{12} = \frac{d_{i1} - d_{i2}}{\sqrt{[(s_{i1} * s_{i1}) + (s_{i2} * s_{i2})]}}$ in which $(S_{i1}^2+S_{i2}^2)^{½}$ estimates the expected standard error of the difference between d_{i1} and d_{i2} (also see Wright and Stone, 1979). The obtained value of t_{12} is compared to the critical value of t. For example, if an alpha level of .05 is used, item i is considered to have significant DIF when $|t_{12}| > 1.96$. Another method commonly used within the 1-PL/ Rasch framework is to examine whether the significant displacement values are found when the item calibrations are anchored (i.e., fixed as expected scores). Typically, item calibrations are anchored using values obtained from the reference group, and displacement values are examined by analyzing the focal group data. The displacement is defined as the "(observed score-expected score)/modeled score variance", which is used to test the hypothesis that the data are generated with the expected scores. An item with a significant displacement value (e.g., > 2 standard error deviation from the expected value) is considered as demonstrating uniform DIF between reference and focal groups. The displacement values can be requested by using WINSTEPS (Linacre, 2005). There are many other software packages that can be used to conduct Rasch or 1-PL IRT analysis and generate item calibrations, e.g., RUMM and ConQuest.

Kubinger (2005) presents situations in which Rasch models are appropriate and discusses approaches to examining model fit. For example, samples can be partitioned into groups, e.g., gender and fit tested using Anderson's likelihood ratio test. The problem arises that with large sample sizes, even with adjustments for multiple comparisons, the type 1 error rate may be inflated (larger than the nominal level), resulting in the spurious classification of items as malfitting (e.g., Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Kubinger makes the important point that magnitude of violations should be considered. Are the differences in parameter estimates of clinical consequence? Graphical displays of such differences can aid in interpretation.

In some of the articles, an extended Rasch model was used, in which the fit between the data and the model is examined using analysis of variance of residuals (ANOVA; Hagquist & Andrich, 2004). Each person is classified according to one of N class intervals, and according to the studied group variable, e.g., gender. A byproduct is a set of residuals that can be used in a two-way ANOVA. The class intervals are chosen to ensure sufficient cell sizes. A significant class-interval effect, irrespective of the group variable, indicates that the item does not fit the model across the construct continuum. A significant group effect, controlling for class interval is an indicator of uniform DIF, while a significant interaction between class-intervals and group is indicative of non-uniform DIF, although the discrimination parameters are still constants.

Analysis of variance using Rasch logits for detection of DIF is an extension of the t-test used to examine differences in the difficulty parameter between groups. A comparison of the Rasch model ANOVA procedure with the Mantel-Haenszel and logistic regression approaches has shown favorable performance for detection of uniform DIF; however, logistic regression showed better performance in the detection of non-uniform DIF (see Whitmore & Schumacker, 1999). Rasch models may not be optimal for the detection of DIF in health data because of the assumption of equal discrimination parameters; however, assuming adequate fit to the data (see Kubinger, 2005; Zumbo and Thomas, 1997), it is often used in circumstances in which sample sizes are smaller (see Lai, Teresi & Gershon, 2005).

The two-parameter logistic and graded response model (Samejima, 1969), used in programs such as IRTLRDIF (www.unc.edu/~dthissen/dl.html) to model ordinal (polytomous) data produces discrimination and difficulty parameters that are the basis for several of the magnitude estimates related to the probability differences (SPD, UPD; Camilli & Shepard, 1994) and Differential Functioning of Items and Tests (DFIT) methodology (Raju, van der Linden & Fleer, 1995). Examples of their use can be found in Morales et al. (2006), Teresi, Kleinman & Ocepek-Welikson (2000) and Teresi et al. (2007). DFIT also accommodates the one parameter (Rasch) model, although there is less experience with applications of this model to DFIT. Several magnitude and impact measures are available in the context of the DFIT methodology.

The Rasch model was used in 10 of the 32 articles included in the table; most with respect to quality of life and general health measures. Four authors (Bjorner, et al. 2004; Chan, Orlando, Ghosh-Dastidar, Duan & Sherbourne, 2004; Hepner, Morales, Hays, Edelen, & Miranda, 2008; Kim, Pilkonis, Frank, Thase & Reynolds, 2002) used an IRT method other than Rasch; these included a two parameter graded response or generalized partial credit model and one author used a non-parametric IRT approach (Moorer, Suurmeijer, Foets & Molenaar, 2001).

A latent variable approach linked to IRT as originally proposed by Birnbaum (Lord & Novick, with contributions by Birnbaum, 1968) is the multiple indicators, multiple cause (MIMIC; Jöreskog and Goldberger, 1975; Muthén & Muthén, 1998-2004) approach to DIF detection (see also Thissen, Steinberg & Wainer, 1993). This approach can be characterized as a confirmatory factor analysis model that incorporates a threshold (difficulty) value. This model uses equality constraints to test whether the item parameters differ between groups. The test of DIF is whether the likelihood of a symptom is different between groups after controlling for disorder as well as other covariates. The DIF measure is the coefficient of the path relating the studied background variable (covariate) to the item, after controlling for the indirect effects (through the latent variable) of other covariates on the item. MIMIC models examine the magnitude of DIF through examination of the direct effect estimate (which detects any residual variance in item response associated with membership in a particular group).

Six studies reviewed in Table 1 used the MIMIC approach. For example, Gallo and colleagues (1998) examined differential performance of depression items between African Americans and Whites by estimating the direct effects of item parameters (DIF indicators), after controlling for the indirect effects of covariates. Yang and Jones (2007) also applied this method to depression items, and Yu and colleagues (2007) used MIMIC to examine items measuring health. MIMIC has been used (Grayson, MacKinnon, Jorm, Creasey & Broe, 2000) to determine the impact of bias on depression scale scores through examination of the bias effect on the total score, estimated as the sum of all the direct loadings from a specific predictor (background characteristic) to the items, contrasting this with the genuine effect that arises from the predictor on the latent variable. Impact has also been examined using MIMIC by comparing the estimated group effects (differences in the coefficients relating group membership to disability) in models with, and without, adjustment for DIF (Fleishman, Spector & Altman, 2002). A major advantage of MIMIC is the inclusion of covariates; possible disadvantages include the inability to examine non-uniform DIF.

Disadvantages of the MIMIC, Rasch and other IRT-based methods are that violations of model assumptions and lack of fit can lead to false DIF detection. All models have to be checked carefully, thus adding to the steps necessary to properly implement the method. An issue in the measurement of self-reported psychological and physical health is the nature of the construct, and whether model misspecification can occur if items are generative rather than emergent. Most latent variable measurement models that form the basis for many DIF detection methods assume that the latent factor causes the symptoms. As an example, it is assumed that indicators of a physical health disorder factor, such as heart disease, blood pressure, shortness of breath, etc. are correlated because they are caused by physical health disorder. However, if the indicators cause the physical health disorder (they are generative rather than emergent), model misspecification may result (Cohen, Cohen, Teresi, Marchi & Velez, 1990; see also Bollen & Lennox, 1991; Fayers & Hand, 1997; Fayers, Hand, Bjordal & Groenvold, 1997). Model assumptions play an important role in DIF detection; lack of model fit and violation of model assumptions can result in false DIF detection. Not all of the articles reviewed provided tests of the assumptions discussed below.

Model Assumptions: Multidimensionality can be mistaken for DIF (Mazor, Hambleton & Clauser, 1998). While some models, e.g., MIMIC and LR can accommodate multidimensional data, most models and applications assume unidimensionality of the underlying trait, and there are additional assumptions associated with specific models. Two major approaches to assessing dimensionality are parametric factor analytic or bifactor models (see Reise, Morizot, Hays, 2007), and the non-parametric methods. In the context of the Rasch model, various fit tests have been used to examine dimensionality.

Model Fit: A contributor to inaccurate DIF detection is lack of model fit (e.g., Bolt, 2002). DIF analyses will be incorrect if, for example, a one-parameter model is selected for DIF detection, when a two or three parameter model would better fit the data (Hambleton, 2006). Numerous fit indices (many distributed as chi-squares) have been investigated; most tests of goodness-of-fit are influenced by combinations of sample size, distributional form, and estimation procedure.

While a detailed discussion is beyond the scope of this review, the issue is: does lack of model infit mean lack of dimensionality? Does it mean DIF? Are all three synonymous? It has been argued (e.g., Teresi, 2006a) that the three concepts, while interrelated are not synonymous, and that model fit, model assumptions and DIF should all be tested separately, and not conflated. As pointed out by Kubinger (2005), lack of model fit does not imply lack of unidimensionality. While some, e.g., Roussos and Stout (1996), have taken the view that DIF implies multidimensionality, McDonald (2000) provides several scenarios for DIF that are not necessarily due to a second nuisance dimension. Borsboom and colleagues (2006, 2002) argue that an alternative cause of DIF might be "relative bias" that might occur if an individual is rating him/herself in relation to others in the setting, for example, members of a football team as contrasted with members of some other team sport. Within groups, the item may perform well, and be related to the measure of the underlying construct, but show DIF due to relative bias or to factors such as poor translation. To return to a point made earlier, it is essential that content experts determine the context in which DIF should be studied, and if possible generate hypotheses about potential DIF that can be tested.

Purification: A major assumption of most methods is that all items in the measure other than the studied item are unbiased. Thus, iterative or two-stage purification is recommended in order to avoid erroneous DIF detection (e.g., Clauser, Mazor & Hambleton, 1993; Holland and Thayer, 1988). Methods of purification vary, but all are based on the notion that items with DIF (except for the studied item for most methods) are removed from the item set used to estimate the disability measure. These DIF-free items form the anchor set.

Interpretation of DIF

Several of the studies examined translations of instruments; these can be affected by lack of conceptual equivalence in different groups. Qualitative analyses are thus important in the determination of reasons for DIF, such as changes in content, format, difficulty of words or sentences, and differences in cultural relevance. Roussos and Stout (1996) recommended a substantive (qualitative) analysis in which DIF hypotheses are generated, and it is decided whether or not unintended "adverse" DIF is present as a secondary factor. Substantive reviewers examine item content, and previously published analyses in order to generate DIF hypotheses. This review is followed by statistical analyses comprised of confirmatory tests of DIF hypotheses. This procedure can be extended to patient-reported outcome measures through use of qualitative methods that include focus groups and cognitive interviews (see Nápoles-Springer, Santoyo-Olsson, O'Brien and Stewart, 2006). This process is rarely performed in practice, and usually after the fact; several of the authors reviewed here provided extensive post-hoc evaluation of the possible reasons for DIF. Examples of good substantive reviews are the articles by Azocar et al., 2001; Gallo et al., 1998; Groenvold, et al., 1995; Kucukdeveci, Sahin, Ataman, Griffiths & Tennant, 2004; Kutlay, Kucukdeveci, Gonul & Tennant, 2003; Pagano & Gotay, 2005 and Prieto et al., 2003.

Methods for review

Selection of patient-reported outcome domains and patient populations

Based on considerations of parsimony, this review was restricted to three domains: depression, quality of life and general health. Collectively, these domains were selected for review because they represent important patient-reported outcomes, contain overlapping item content, and are included in major item banking projects, such as the Patient Reported Outcomes Measurement Information System (PROMIS; www.NIHPROMIS.org) project (Reeve, 2006; Reeve et al., 2007). PROMIS, part of the U.S. National Institutes of Health (NIH) roadmap initiative (RFA-RM-04-011), aims to provide an infrastructure for clinicians and researchers by establishing generic item banks across various disease groups, and applications of computerized adaptive testing.

Patient populations were limited to adults. Measures developed for use with children and adolescents were excluded. This decision was made in the interest of parsimony, and because measurement in children encompasses different issues; constructs may not be conceptually equivalent, and targeted outcomes could be different for children as contrasted with adults.

Selection of articles and measures for review

Identification of manuscripts addressing measurement equivalence using DIF was conducted through the use of several search engines. An initial search using the Columbia University Library PubMed database was conducted on November 11, 2005. Two parameters were specified: time frame (from 1995 to 2005) and key words appearing in the citation and the abstract, i.e., “Differential Item Functioning.” An Ovid-assisted cross-referencing search was conducted on the same date, using the same parameters. After deleting duplicates a total count of 120 unique references were identified. A web-based ProQuest search conducted on December 2, 2005, as a second crosscheck, failed to identify any new articles. An additional search was conducted using the Columbia University Library PubMed database on February 6, 2006 expanding the dates to include 2006 and the key words to include item bias. A search was conducted on February 12, 2006 through the Northwestern University library system using the keywords, “quality of life,” “health-related quality of life,” “DIF,” and “item bias.” Finally, a search was conducted in July, 2008. The purpose was to check the results of the search in the areas of quality of life, depression and general health; several additional articles were identified.

The selected articles from the first two searches were then divided (in terms of their content) into “methodological” and “applied”. Only applied articles (i.e., articles in which DIF was applied to a specific measure) were retained for evaluation. However, methodological articles were used in the overall review. A second-level iteration of manuscript selection was then conducted in which only manuscripts focusing on measures of the constructs of interest (depression, general quality of life and general health) that used samples from adult populations of sufficient size were included.

Measurement Evaluation Grids (MEGS)

Table 1 is a summary of some of the information contained within the MEGS, which were created in order to provide evaluative criteria for measurement review. The MEGS contain elements determined by the measurement cores of the U.S. Resource Centers for Minority Aging Research (RCMAR) to be important in the evaluation of measures for equivalence across groups differing in characteristics such as ethnicity, literacy and education. Additional information can be found at the following website (www.research-HHAR.org). Elements include: sample characteristics (recruitment, data collection methods, response rate), format of the measure (design, readability, type (level) of measurement, scoring (range, direction, rules and missing data), translations, psychometric properties (scale construction, basic summary statistics, variability, test-retest, interrater, internal consistency reliability, content, construct, concurrent and predictive validity, sensitivity to change) and differential item functioning (variables studied, sample size, DIF method used, tests of model assumptions, purification, evidence of uniform and non-uniform DIF, magnitude and impact of DIF), and review of strengths and weaknesses. Specific comments related to each article appear in Table 1, in the column titled, "Review".

Guidelines for review

Presented above are brief definitions of the elements used in the MEGS that are relevant to examination of DIF. These include methods for examination of DIF, tests of model fit and assumptions, purification, evidence of uniform and non-uniform DIF, magnitude and impact of DIF, all of which can affect the detection rate (e.g., Rogers & Swaminathan, 1993; Whitmore & Schumacker, 1999). Note that basic psychometric analyses such as reliability and validity of the measures are not the focus of this review, but are assumed to have been examined.

Results: Review of measures

Shown in the table are the results of DIF analyses for the measures reviewed. For brevity, only selected elements from longer internal summary reviews of each article are shown. These include the name of the measure, source of the DIF analyses, the method used, the results and the summary of the methodological review. Because this information is included in the table, it is not repeated; rather each section below is a brief summary of findings and recommendations for future work.

Depression

Two forms of severe mental illness, unipolar major depression and bipolar disorders, have been identified by a World Health Organization study (Lopez & Murray, 1998) to be among the leading causes of disability. Recent findings from the National Health Interview Survey (Pratt, Dey & Cohen, 2007) showed that the prevalence of serious psychological distress was more than twice as high (5.9%) among Hispanic older persons (65 and over) than among non-Hispanic Blacks (2.4%) and Whites (2.1%). These differences among race and ethnic groups were not observed among younger cohorts. Several studies (e.g. Callahan & Wolinsky, 1994; Gallo et al., 1998; Koenig et al., 1992; Teresi et al., 2002) of depression have shown lower rates of depression among Blacks as contrasted with other groups, and among older as contrasted with younger cohorts. In order to determine if differences in rates between race/ethnic, age and gender groups reflect actual differences and not item bias, studies of factorial invariance and DIF are needed.

Several constructs or appellations related to depression include dysthymia, emotional distress, general distress, serious psychological distress, affective disorder and anxiety. Typically a measure labeled as one of the above contains at least some depression items, and some items from most depression scales are also found on measures of these other constructs. There is a vast literature discussing the interrelationships among some of these constructs (e.g., Hockwarter, Harrison & Amason, 1996; Huelsman, Nemanick & Munz, 1998; Lawton, Kleban, Rajagopal, Dean & Parmelee, 1992; Teresi, Abrams & Holmes, 2000), a discussion of which is beyond the scope of this presentation. However, because the majority of the measures reviewed below are intended to measure depression, this is the term that is used here to discuss the constellation of symptoms examined in terms of factorial invariance and DIF.

Several studies have examined the factor structure of depression measures, most using exploratory factor analyses; however, recently studies using confirmatory multi-group factor analyses have emerged. Strict residual-level factorial invariance (using a multi-group factor model) is equivalent to DIF testing using a 2-parameter IRT model (see Meredith and Teresi [2006] for a discussion). Most of the 14 DIF studies of depression used a variant of IRT; six used the one parameter Rasch or other IRT models, including the two parameter model with likelihood tests, and five used MIMIC or restricted factor analyses. Others used the Mantel-Haenszel ordinal logistic regression or SIBTEST (Shealy and Stout, 1993) methods.

An important first step in the conduct of DIF analyses is examination of dimensional invariance because unidimensionality is an assumption of most methods used. Dimensionality is usually tested using a factor analytic approach. Reviewed below are studies of factorial invariance or DIF in measures of depression; another review of two popular depression measures can be found in Mui, Burnette, and Chen (2001). The reviews related to depression presented in the MEGS include the years 1995 to 2008; however, some important earlier work is briefly reviewed (for a more detailed review of these earlier studies, see Teresi & Holmes, 1994, 2001)

Centers for Epidemiological Studies Depression (CES-D; Radloff, 1977): Studies of factorial invariance and individual factor analyses with different groups have been conducted with respect to various measures of depression. For example, Foley, Reed, Mutran, and DeVellis (2002) examined the factor structure of the CES-D among older African Americans using an exploratory factor model. The eigenvalue for the first factor was 6.26, explaining 31% of the variance, while the second was 1.63, explaining 8% of the variance; thus an essentially unidimensional depression/somatic symptoms factor characterized the data. The results demonstrated that the factor structure was different from that observed in previous studies, and that collectively across studies, divergence among solutions was observed. Gregorich (2006) subjected the Somatic and Retarded Activity factor of the CES-D (originally reported by Radloff using exploratory factor analyses) to confirmatory factor analyses. Samples of Black and White men over age 50 were studied; the results indicated that metric invariance was achieved for all five items; however, the item, ‘effort’ did not achieve strong factorial invariance of item intercepts, and ‘appetite’ was not strictly invariant (equal residual variances).

The majority of recent studies of DIF in depression measures have focused on the CES-D. The group variables as well as the method used to examine DIF varied across studies, however. Using an IRT log likelihood approach, Chan et al. (2004) found 12 items manifesting DIF (uniform and/or non-uniform): “happy”, “enjoyed life”, “could not get going”, “talked less than usual”, “everything was an effort”, “felt as good as others”, “felt depressed”, “felt sad”, “trouble keeping my mind on what I was doing”, “people dislike me”, “my life had been a failure”, and “felt hopeful about the future”, when mode effect (phone vs. mail) was examined. Discussing the impact of DIF, the authors noted that DIF could result in an increase of up to six points on the depression continuum for the mail respondents. On the other hand, Cole et al. (2000) found 17 of the 20 CES-D items to be relatively free of item bias by age, gender, and racial groups. Only three items, “people are unfriendly”, “people dislike me”, and “crying spells” were found to function (uniformly) differently among subgroups of gender and race. The magnitude of DIF on the interpersonal items, however, reflected proportional odds up to three points higher for Blacks as compared to Whites with equivalent levels of depressive symptomatology. Similarly, the magnitude of DIF on the “crying spells” item showed a two-point increase in proportional odds for women as compared to men, matched on overall depressive symptoms. These artificially increased odds for endorsing such items by Blacks and/or by women could carry as an overall bias at the scale level. The authors highlighted a shorter, relatively DIF-free version of the CES-D, which correlated .99 with the original scale. More recently, Yang and Jones (2007) replicated the findings of Cole and colleagues (2000), using a latent variable model approach, MIMIC. Data from the New Haven Established Population for the Epidemiologic Studies of the Elderly were used to examine DIF related to age (75 and over vs. younger), gender, and race (Black vs. White). Blacks were more likely to respond in a higher category, conditional on depression to the items: “people are unfriendly” and “people dislike me”. The proportional odds for women were higher than for men for the item “crying spells”. These items had relatively large magnitude of DIF because the proportional odds for these items were all two or above. Using Rasch analysis, similar findings were reported by Covic, Pallant, Conaghan and Tennant (2007). “I felt tearful” and “I had crying spells” had significant DIF on both age and gender. The subgroup aged 53 years or less (compared to 54-65 & 66+) and females (compared to males) were significantly more likely to endorse these two items.

In their examination of the contribution of specific physical disorders to uniform DIF in the CES-D, Grayson et al. (2000) found item-specific effects for age, gender, and marital status.

> *Older participants reported being more “bothered by things” and less “hopeful about the future”. Men found things “less of an effort”, were “less fearful”, “slept better”, and reported “crying less”; being widowed was associated with “feeling at least as good as others”, and with more “fear and loneliness”. (pg 276)*

Additionally, mobility, ADL, and IADL impairment showed direct effects on “poor appetite”, “finding everything an effort”, “restless sleep”, and “inability to get going”. In terms of physical disorders, heart disease, stroke, any other systemic disease, gait instability, and cognitive impairment showed positive association with depression. However, individuals with physical disorders, for reasons unassociated with depression, underreported on items such as: “felt as good as others”, “talked less than usual”, “people are unfriendly”, “enjoyed life”, “crying spells”, “felt sad”, and “people dislike me”, and showed higher endorsement on items such as: “poor appetite”, “everything an effort”, and “inability to get going”. As discussed by the authors, depending upon the group variable being examined, the impact of DIF on the total CES-D score ranged from trivial to considerable (over seven times the magnitude of the effect on depression). Similarly, Pickard, Dalal, and Bushnell (2006) found that four items, “My sleep was restless,” “I felt that people disliked me,” “I did not feel like eating,” and “I had crying spells”, demonstrated statistically significant uniform DIF when stroke and primary-care groups were compared. The authors do not recommend stroke-specific modifications to the CES-D, however, arguing that only one item was identified as uniquely psychometrically problematic. It is noted that the small sample size for the stroke patient group renders the study results exploratory.

Gelin and Zumbo (2003) examined the CES-D using the Health and Health Care Survey data collected from 600 community resident adults residing in Northern British Columbia, Canada (290 women and 310 men). Using an ordinal logistic regression approach, they found that the manner in which items were scored affected DIF results, as did the endorsement proportions. Depending if items were scored as binary, ordinally, or according to a persistency (frequency of at least 3 to 7 days) threshold, results changed. The “crying” item showed high magnitude of gender DIF for both binary and ordinal scoring methods, in the direction that the conditional endorsement was higher for women than for men; the item was a much more severe indicator for men because it takes higher levels of depression before men will endorse this item. While DIF was observed for two other items (“effort” and “hopeful”) using the persistence method, the low item prevalence renders these results less robust.

General Health Questionnaire (GHQ; Goldberg, 1972): Several factorial invariance studies of the General Health Questionnaire-12 have been performed. This 12-item measure assesses minor psychiatric disorders and has been viewed as a measure of general distress. Items include “concentration,” “sleep disorder due to worry,” ”feeling depressed,” “worthless,” “unhappy,” and “lack of enjoyment of activities.” Three factors have been observed: anxiety/depression, social dysfunction, loss of confidence (Shelvin and Adamson, 2005). However, these analyses showed that a higher order factor or a 12 item summary measure may be sufficient; factorial invariance of loadings, error variance and factor variances was established for gender. The authors of another study (Makikangas et al., 2006) of the GHQ-12 (ignoring minor misfit and one invariant factor loading) demonstrated factorial invariance of thresholds, loadings and factor means over time. Jorm et al. (2005) examined the factorial invariance of the Goldberg Depression and Anxiety scales using a community survey of 7485 persons in several age categories: 20-24, 40-44, 60-64. These authors established weak (metric) factorial invariance for the two factors. A generalized measure of psychological distress was also recommended: the sum of all items, excluding one, “difficulty falling asleep”. A potential weakness, acknowledged by the authors is that older age groups were not examined. Additionally, only metric invariance was established. Duncan-Jones, Grayson and Moran (1986) used IRT to examine the gender bias of the 12-item GHQ. Two items (“feeling constantly under strain” and “feeling unable to overcome difficulties”) were more related to depression for women than they were for men.

SHORT-Comprehensive Assessment and Referral Evaluation (CARE; Gurland, Golden, Teresi & Challop, 1984): An examination of item bias (using IRT) associated with the SHORT-CARE Depression scale, was conducted by Teresi and Golden (1994); these authors found that some of the somatic symptoms (“headaches”, “crying”, and “lack of interest”) were relatively less severe indicators of depression for Latinos than for White, non-Latinos. “Crying” was of higher DIF magnitude. Across the disability spectrum, the likelihood of endorsement of this item in particular was higher for Latinos than for White, non-Latinos. A subset (with some modifications) of the SHORT-CARE Depression Scale, including the “crying” item, is contained within the EURO-D, a widely used measure that has recently been evaluated psychometrically (Castro-Costa, Dewey, Stewart, Banerjee, Huppert, Mendonca-Lima, et al., 2008).

Beck Depression Inventory (BDI; Beck, Ward, Mendelsohn, Mock, and Erbaugh, 1961): Gibbons, Clark, Vonammon-Cavanaugh, and Davis (1985) used IRT to examine the BDI, comparing medically ill inpatients with psychiatric patients. The vegetative symptoms, “loss of weight” and “of sexual interest” were particularly poor discriminators of depression severity among the medically ill sample. Two items (“loss of satisfaction”, “loss of social interest”) were found to maximally assess depression severity. Azocar et al. (2001), examining uniform DIF identified four BDI items to be biased for the Spanish- (vs. English) speaking sample. The items, “I feel like I am being punished”, “I feel like crying”, and “I believe I look ugly”, were more likely to be endorsed, and the item “I can’t do any work at all” was less likely to be endorsed by Spanish speakers regardless of their level of depression. The authors point out that the impact of DIF in this scale is such that it could result in an artificial increase of the mean scores for Latino samples up to six points (possible scores ranged from 0 to 30) above those of English-speaking samples with equivalent depression levels. Kim et al. (2002) using item response theory to examine the contribution of age to DIF in the BDI, found three items reflecting uniform DIF across all levels of depression: “loss of libido”, “weight loss”, and “disappointment in self’, in which midlife patients were more likely to endorse “loss of libido” and “disappointment in self”, and less likely to endorse “weight loss” than late-life patients. This finding is similar to the earlier study reviewed above (Gibbons, et al., 1985) that found that the vegetative symptoms (“loss of weight” and of “sexual interest”) were not well related to depression severity among a medically ill sample. Kim et al. also found non-uniform DIF on 8 of 11 BDI items: “self-criticism”, “social withdrawal”, “irritability”, “guilt feelings”, “sense of failure”, “sleep disturbance”, “somatic preoccupation”, and “work inhibition”. They found that the impact of DIF on the BDI was not trivial, given that approximately half of the items on the scale accounted for 80% of the differential test functioning.

Geriatric Depression Scale (GDS; Yesavage, et al., 1982): In contrast to the above findings of DIF related to several items in the BDI, Tang, Wong, Chiu, Lum, and Ungvari (2005) failed to document any uniform or non-uniform DIF of the GDS with respect to age, education or cognitive impairment in a sample of Chinese patients. However, more recently, Broekman, Nyunt, Niti, Jin, Ko, Kumar, et al. (2008), examining DIF in a heterogeneous Asian population, found ten of the GDS-15 items to show DIF associated with age, gender, ethnicity and chronic illness. Specifically, six items, e.g., “drop many activities and interests”, “prefer staying home”, “more problems with memory”, “feel pretty worthless”, “not full of energy”, and “not happy most of the time” showed age-related DIF. Five items showed gender-related DIF: “afraid that something bad is going to happen”, “prefer staying home”, “more problems with memory”, “feel situation is hopeless”, and “not satisfied with life”. The four items that showed ethnicity-related DIF were “prefer staying home”, “think not wonderful to be alive”, “feel pretty worthless”, and “more problems with memory”. Finally, two items showed illness-related DIF: “feel pretty worthless” and “not full of energy”. The authors concluded that the cumulated effects of specific item bias due to age, gender, ethnicity and chronic illness could potentially bias the total test score.

Depression Diagnostic Scales: Items from depression diagnostic scales were also investigated for DIF. For example, Gallo et al. (1998), examining the Diagnostic Interview Study (DIS; Robins, Helzer, Croughan & Ratcliff, 1981) for racial bias, found certain items showing uniform DIF. "Sleep disturbance" and "sadness" were less likely to be endorsed by African Americans, and "difficulty with concentrating" and "thoughts of death" were more likely to be reported by older African Americans than by Whites. No discussion of the impact of DIF was presented, however.

Other research examining DIF associated with Diagnostic Statistical Manual (DSM) symptoms has been performed (Simon and Von Korff, 2006). The Composite International Diagnostic Interview (CIDI) (Kessler, Wittchen, Abelson, McGonagle, Schwarz and Kendler, 1998), a DSM III-R-based diagnostic interview schedule, showed race/ethnicity-related DIF (Breslau, Javaras, Blacker, Murphy, and Normand, 2008). Blacks (vs. Whites) were less likely to endorse "lack of energy", "felt worthless" and "thoughts of suicide" at the item level, and "loss of energy" and "self-reproach" at the symptom level. Similarly, underestimation of depression was found among Hispanics (in contrast with Whites) for "increased weight" and "waking early" at the item level and for "suicidality" at the symptom level. In another study, the Patient Health Questionnaire depression scale (PHQ-9; Kroenke, Spitzer, and Williams, 2001), a criterion-based depression measure based on symptoms from the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), was examined. Exploratory factor analyses were performed across four samples: African American (n=598), Chinese-American (n=941), Latino (n=974) and non-Hispanic Whites (n=2,520) yielded essentially unidimensional factors across samples (Huang, Chung, Kroenke, Delucchi, and Spitzer, 2006). Eigenvalues ranged from 3.5 to 4.42, and variances explained were: 38.9 (Chinese), 39.6 (Latino), 40.1 (African American), and 49.1 (non-Latino White). Testing for DIF, items such as "anhedonia", "sleep", and "appetite" showed significant DIF using the Mantel-Haenszel statistic in the Chinese-American group and Latino groups as compared to White non-Latinos. In addition, "depressed mood", "low energy", "appetite change", and "low self-esteem" evidenced DIF for Latino as contrasted with non-Latino Whites. Most of the DIF found at the item level in those two groups, using the Mantel-Haenszel method no longer showed a significant level of DIF after controlling for covariates of age, gender, and English-language ability in the MIMIC model test. No item-level DIF was reported for the African American group. While there were few scale group differences before DIF adjustments among the groups defined by race/ethnicity, there were some differences in proportion over threshold classified as depressed. Additionally, subgroup comparisons involving gender, age and language could be influenced by DIF. 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, Williams, Kroenke, Linzer, deGruy, Hahn, et al., 1994), Hepner and her colleagues (2008) did not find any significant DIF among lower income Black and White women.

Summary: The response patterns and the factorial composition of scales assessing depressive symptomatology have been found to be affected by several factors (Mui et al., 2001; Pedersen, Pallay, & Rudolph, 2002). For example, the affective CES-D item tapping sadness showed DIF based on physical disorder and interview mode, and a similar DIS item reflected DIF based on race. Additionally, the CES-D interpersonal items "people are unfriendly" and "people dislike me" showed DIF with respect to one or more of several variables: interview mode (Chan et al., 2004), gender, race (Cole et al., 2000), physical disorder (Grayson et al., 2000), and stroke (Pickard, et al., 2006). DIF was also observed in the "crying" item of the CES-D by age, gender, race, physical disorder, and stroke condition (Azocar, et al., 2001; Cole, 2000; Covic, et al., 2007; Gelin and Zumbo, 2003; Grayson et al., 2000; Pickard et al., 2006; Reeve, 2000; Teresi and Golden, 1994; Steinberg and Thissen, 2006; Yang and Jones, 2007).

Impact of DIF on depression measures: The impact of DIF in the CES-D, as discussed by the respective authors, ranged from trivial to significant, depending on the reference group studied. For example, in some studies DIF was found to result in a considerable artificial increase in the overall depression score for mail responders (Chan et al., 2004), and for Blacks (Cole et al., 2000); on the other hand, scale adjustments were not warranted for stroke patients (Pickard et al., 2006). Similarly, Osborne and colleagues (2004), discussing the impact of DIF in the HADS (Zigmond & Snaith, 1983), did not recommend adjustments for cancer patients. The impact of DIF in the BDI (Beck et al., 1961) was demonstrated to be sizable, showing artificial, favorable endorsement of some of the items by Spanish-speaking Latinos (in contrast to English speakers; Azocar et al., 2001); similarly, another analysis demonstrated that half of the items in the scale accounted for 80% of the differential test functioning (Kim et al., 2002).

In summary, about two thirds of the studies reviewed in the area of depression examined magnitude, and almost all estimated the impact of DIF; a little over one third examined non-uniform DIF. In general, findings were of large amounts of DIF of sizeable magnitude and impact. Adjustments of scale scores were frequently recommended.

Quality of life

Quality of life has been conceptualized in many different ways (see Lawton, 1991; Katz and Gurland, 1991; Gurland & Gurland, in press). It is beyond the scope of this article to discuss issues related to the definition, except to point out that overlapping domains (e.g., emotional, physical and functional states) are included in most generic quality of life measures. It is also noted that a distinction can be made between health-related quality of life and general quality of life; the latter includes additional dimensions such as environmental and personal resources (Albert and Teresi, 2002). Given that most of the quality of life measures reviewed here focus on physical, functional and emotional states, most of the findings generalize to DIF and health-related quality of life.

Various approaches were used to examine DIF in the measures of health-related quality of life; most feed into the IRT framework. While parametric IRT examines DIF at the item level, some non-parametric IRT methods (e.g., Mokken) define DIF at the response category level by testing the hypothesis of "equal item step order" across subgroups. Among the nine studies of DIF in quality of life measures reviewed, four used the Rasch model; logistic regression was used in two studies; the two-parameter model was not used in these articles.

Generic emotional distress measures: Instruments that measure quality of life often contain items related to emotional health. A recent analysis (Crane, Gibbons, Ocepek-Welikson, et al., 2007; Teresi, Ocepek-Welikson, Kleinman, Cook, Crane, Gibbons, et al., 2007) of 15 emotional distress items from a study of quality of life included several depression items from a number of scales (Cancer Rehabilitation Evaluation System [CARES-SF; Ganz, Schag, Lee & Sim, 1992; Schag, Ganz, & Heinrich, 1991]; European Organization for Research and Treatment of Cancer Quality of Life Questionnaire [EORTC; Aaronson, et al., 1993]; Functional Assessment of Cancer Therapy [FACT; Cella, 1997], Medical Outcomes Study Short Form Health Survey [SF-36; Hays, Sherbourne, & Mazel, 1993; McHorney, Ware & Raczek, 1993; Ware and Sherbourne, 1992]). The findings from two analyses of these data showed that several items evidenced uniform DIF. Conditional on positive mental health, White respondents were more likely than African Americans to report that they were "not worried about dying" and "did not feel worried" (Teresi, et al., 2007). Crane, Gibbons, Ocepek-Welikson, et al. (2007) also identified "feeling worried" as showing DIF for the comparison of African Americans and Whites; however, this was not significant after Bonferroni correction. Women were more likely to say that they were "able to enjoy life" and were "content with their quality of life" (Teresi et al.); Crane, Gibbons, Ocepek-Welikson, et al. (2007) also identified "content with quality of life" as showing gender DIF before, but not after Bonferroni correction. Older respondents (66 and over) as contrasted with the younger cohort were more likely to report that they "were not worried about dying"; "feeling calm and peaceful" also showed age DIF (Teresi, et al., 2007). Crane, Gibbons, Ocepek-Welikson, et al. (2007) did not find DIF related to age, but did find one item ("being a happy person") to show DIF for marital status. These authors concluded that there could be DIF impact associated with race on the General Distress scale for some individuals.

European Organization for Research & Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30; Aaronson, et al., 1993) – *emotional health items:* Several authors have examined DIF in the EORTC QLQ-C30 emotional function subscale. Two items repeatedly showed uniform DIF in studies of different languages and ethnic groups, "Did you worry" (in six studies) and "Did you feel irritable" (in three studies). In a study that compared language groups (Bjorner, et al., 2004), 'worry' was estimated to be the most informative, with the largest mean threshold parameter, and one of the largest slopes. In a study of 359 Caucasian, Filipino, Hawaiian, and Japanese cancer patients, the 'worry' item was significantly less difficult for Hawaiians (Pagano & Gotay, 2005). The item, "Did you worry" showed DIF for English vs. Norwegian, Dutch, and French speakers (Petersen et al., 2003) in a cross-national study of 10 countries. This item also showed non-uniform DIF when comparing subgroups of Norwegian with English speakers. A review of 13 translations of the EORTC performed by Scott, Fayers, Bottomly, Aaronson, de Graff, Groenvold, et al. (2006) found that respondents using the Norwegian, Turkish or the two Chinese translations were less likely to endorse "Did you worry" compared to English speakers, while Germans were more likely to endorse this item. An examination of several EORTC items by Teresi and colleagues (2007) found significant yet moderate DIF in the direction that Blacks required more positive health in order to endorse this item. In a review by Bjorner et al. (2004), "Did you feel irritable?" consistently had the smallest slope and mean threshold, indicating that respondents were less likely to report symptoms based on this item, and it provided markedly less information than the other items. In the ten-country cross-national study, "Did you feel irritable" showed DIF for English vs. Norwegian, Spanish, and German speakers (Petersen, et al., 2003). Scott et al. (2006a) also found DIF for Spanish speakers, and in addition, found DIF for Dutch speakers. This item also showed non-uniform DIF for subgroups of Spanish and English speakers (Bjorner). Two separate studies found two additional items in the subscale and showed DIF for language groups: "Did you feel tense" for English vs. Swedish and Spanish speakers (Petersen, et al., 2003) and English vs. Polish and

Singapore Chinese speakers (Scott, et al., 2006a). The second item, "Did you feel depressed" showed uniform DIF for English vs. Norwegian and Swedish speakers in both studies. In addition, German, and Finnish speakers (Petersen et al., 2003) and Polish and the two Chinese translations (Scott, et al., 2006a) showed uniform DIF for this item.

EORTC – physical function items: In a study that examined item bias in the EORTC QLQ-C30 (Aaronson et al., 1993) among age groups and form of treatment among Danish breast cancer patients (Groenvold et al., 1995), one physical function item, "Do you have to stay in a bed or a chair for most of the day?" was biased across both age and treatment groups. This item was also found to show DIF for several language and cultural groups (Scott, et al., 2006a, 2007). In addition, "trouble doing strenuous activities", "taking a long walk" and "taking a short walk" showed DIF among age (Teresi, et al., 2007), language (Scott, et al., 2006a), and the latter two for cultural groups (Scott, et al., 2007). Crane, Gibbons, Ocepek-Welikson, et al. (2007) also reported uniform DIF that remained after Bonferonni correction for age for "trouble with a long walk", and "trouble doing strenuous exercise" for gender.

Scott and colleagues found that Turkish vs. English speakers and Islamic (Turkey, Iran and Egypt) vs. UK groups required more "help with eating, dressing, washing or using the toilet" (Scott, et al., 2006a, 2007). Among Turkish respondents, the item, "I find it difficult to take care of people I am close to", showed DIF for age. Younger, in contrast to older respondents, were less likely to report difficulty in caring for persons to whom they are close (Kutlay et al., 2003).

EORTC – additional items: The item, "Did pain interfere with your daily activities", showed DIF across language (Scott, et al., 2006a), cultural groups (Scott, et al., 2007), and cancer treatment groups (Groenvold, et al., 1995). Those receiving chemotherapy were less likely to report "difficulty remembering things". Relative to other language comparison groups, English speakers scored significantly lower on the item, "did pain interfere with your daily activities" (Scott et al., 2006a). South Western Europeans were significantly less likely to endorse "have you had pain" (Scott et al., 2007). In a separate study, Caucasians showed significantly less difficulty with "work at job" and "constipation". Caucasians and Japanese had greater difficulty with "social activities" than did Hawaiians and Filipinos (Pagano & Gotay, 2005). Danish and German speaking respondents were more likely to endorse the "family life" item, and along with the Spanish speaking group, score lower on the "social activities" item than did English speakers (Scott et al., 2006a).

EUROQoL – emotional and physical function items: A Rasch analysis of the EUROQoL (Kind, 1996) performed by Prieto and colleagues (2003), identified the emotional function item, 'Anxiety/ depression' as the easiest (more likely) item to endorse across ten European countries. 'Mobility' and 'self-care' were the most difficult (least likely) items to endorse for all countries except Denmark, where respondents were more likely to endorse the 'mobility' item (Prieto et al., 2003).

Other quality of life measures: A study of the Turkish translation of the Rheumatoid Arthritis Quality of Life Scale (De Jong, Van der Heijde, McKenna & Whalley, 1997) found two biased physical function items, "I find it difficult to walk to the shops", and "I sometimes have problems using the toilet". The latter was more difficult for Turkish respondents than Americans. The authors also found uniform cross-cultural DIF in the items "Often gets frustrated" and "Feels unable to control situation".

An examination of the WHOQOL-BREF found DIF for age groups for “able to get around”, “satisfied with sex life”, and “ability to get things you like to eat”. In addition, four items exhibited DIF between elementary, secondary and higher education groups (Wang, Yao, Tsai, Wang, & Hseih, 2006).

Impact of DIF on quality of life measures: Groenvold and colleagues (1995) and Pagano and Gotay (2005) examined the impact of removing items with DIF from the scales; both concluded that biased items should not be removed. Bjorner and colleagues (2004) examined the effect of removing items with DIF from the scoring algorithm; they concluded that scoring algorithms that took language-related DIF into account did not perform as well as those that ignored DIF. In contrast, an analysis performed by Petersen et al. (2003) found that scale scores were equivalent when the biased item, “worry” was removed from the scale. No recommendations were made to remove items.

General health

The general health measures often contain subscales or items measuring several domains, including physical and mental health, pain, vitality and social role functioning. Five authors of the nine studies in this area examined a commonly used instrument to measure general health, the Short Form (SF)-12 (Ware, Kosinski & Keller, 1996), SF-36 (Ware and Sherbourne, 1992; Ware, Gandek & The IQOLA Project Group, 1994) or RAND-36 (Hays, Sherbourne, & Mazel, 1993). A little over half of the nine studies of general health applied a latent variable DIF model, such as Rasch or MIMIC. However, Bjorner and colleagues tested for uniform and non-uniform DIF in the SF-36 using a partial gamma coefficient in Danish and American populations (Bjorner, Kreiner, Ware, Damsgaard & Bech, 1998). Using this method, they found four items from the Physical Functioning and two from the General Health Scale behaved differently for these cultural groups. Fleishman and Lawrence (2003) examined DIF in the SF-12 (Ware et al., 1996) by race/ethnicity, age, gender, and education level for an American, adult sample (aged 17 or older) using the MIMIC model. Ten of the twelve items showed DIF for the demographic variables tested. Yu and colleagues (2007) examined DIF in the physical function (PF) and mental health (MH) subscales of the SF-36 for demographic characteristics and hypertension, rheumatic conditions, diabetes, respiratory diseases, and depression also using the MIMIC model. Uniform DIF was observed for numerous items. Perkins, Strump, Monahan and McHorney (2006) examined DIF in all subscales of the SF-36 in two large national datasets; the National Survey of Functional Health Status contained general population data, and the Medical Outcomes Study contained data from a chronically ill population. Data were examined with respect to age, education, race and gender using proportional-odds logistic regression. While numerous items exhibited DIF, most were not of high magnitude. Moorer and colleagues (2001) examined the RAND-36 (Hays et al., 1993) using a non-parametric IRT model, Mokken scale analysis for polychotomous items (MSP). Using this method, the authors found no evidence of DIF across disease groups (multiple sclerosis, rheumatism, and COPD). In addition, two groups examined DIF in 15 RAND-36 items using the item response theory log-likelihood ratio and ordinal logistic regression approaches in a sample with cancer or HIV/AIDS (Teresi et al., 2007; Crane, Gibbons, Ocepek-Welikson, et al., 2007). High magnitude of DIF was observed for three of the five items identified with uniform DIF.

Short Form measures – physical function items: The majority of DIF findings relate to the physical function subscale. Nine of the ten items exhibited DIF in at least one group; “vigorous activities” was most frequently cited as problematic while “walk one block” showed no DIF findings. “Vigorous activities” showed both uniform and non-uniform DIF for age (Perkins, et al., 2006; Yu, Yu, & Ahn, 2007; Teresi, et al., 2007), education (Perkins, et al.; Yu, et al.), income (Yu, et al.), disease group (Yu, et al.), gender (Perkins, et al.), and race (Perkins, et al.; Teresi, et al.). While older people with poor physical function reported less limitation than younger people, among those with high physical function, older persons reported more limitation than did younger people with “vigorous activities” (Perkins, et al., 2006). A MIMIC model showed negative effects for age for this item (Yu, et al., 2007). Differences of large magnitude were found in sample of cancer, HIV and AIDS patients; conditional on functional status, White respondents were less likely than Blacks, and those 66 and older were less likely than those younger to report that they were capable of “vigorous activities” (Teresi, et al., 2007). In general and sick populations, those with less education and Blacks reported less limitation with respect to “vigorous activities” with more pronounced difference at lower to mid-levels of physical functioning (Perkins, et al.). In addition to similar education findings, Yu and colleagues found that lower income had positive DIF effects (2007). The analyses by Perkins and colleagues of a sample of chronically ill persons, showed non-uniform gender DIF for this item. Finally, in a large data set, those with respiratory disease scored lower than expected on “vigorous activities” (Yu, et al., 2007).

Older people reported more limitations in “moderate activities” (Yu, et al., 2007; Fleishman & Lawrence, 2003), and people with hypertension more frequently endorsed “health limits moderate activities” (Yu, et al.). Conditional on function, fewer Americans than Danes (Bjorner, et al., 1998), Blacks as contrasted with Whites (Teresi, et al.), and females had greater limitations “lifting or carrying groceries” (Yu, et al.; Teresi, et al.). Older people, those with less education, and females reported more limitations with “stair climbing” (Fleishman & Lawrence; Yu, et al.). Age showed both uniform negative effects for “bending/kneeling/stooping” (Yu, et al.), and non-uniform effects (Perkins, et al.; Teresi, et al.). Older people with poor physical function reported less limitation than did younger people. Among those with high physical function, older people reported more limitations than did younger persons with this item. Those with hypertension more frequently endorsed limitations in “bending/ kneeling/ stooping” (Yu, et al.). Crane, Gibbons, Ocepek-Welikson, et al. (2007) found DIF for this item across three forms of logistic regression.

Both uniform and non-uniform DIF were found for “walking more than a mile”. Both Crane, Gibbons, Ocepek-Welikson, et al. (2007), and Teresi and colleagues found DIF in this item using the same data set, but two DIF detection methods (LR and IRT Log likelihood). Whites compared with Blacks, and older compared with younger persons reported more limitations (Teresi, et al.). Among those with low physical functioning, Blacks reported less limitations than Whites “walking more than a mile”, while at higher function levels Blacks reported more limitation than Whites (Perkins, et al.). Age showed negative effects for this item (Yu, et al.). Fewer Danes than Americans indicated limitations in “walking more than a mile” and “walking several blocks” (Bjorner, et al.). Older people (Perkins, et al.), females (Yu, et al.), and Americans (compared to Danes; Bjorner, et al.) reported fewer limitations with respect to “bathing or dressing”. This item also showed non-uniform DIF for age and education (Perkins, et al.).

Short Form measures – mental health items: There were DIF findings for each mental health item in at least one of the three studies in which DIF was evidenced for this subscale. Americans (compared with Danes; Bjorner, et al., 1998) and Blacks (compared with Whites) were more likely to endorse 'been a very nervous person', while those with less than a college education and those who were married were less likely to endorse this item (Yu, et al., 2007). The group with less education showed a negative effect for "nothing could cheer you up" (Yu, et al.). Americans were more likely than Danes to endorse this item (Bjorner, et al.). "Felt calm and peaceful" showed multiple effects across studies. Older age groups (Perkins, et al., 2006; Fleishman & Lawrence, 2003; Yu, et al.; Teresi, et al., 2007), those with low education (Perkins, et al.; Fleishman & Lawrence) and minorities (Fleishman & Lawrence; Yu, et al.) were more likely than expected to give higher ratings. Those who were married endorsed this item less often than expected (Yu, et al.). Fleishman and Lawrence found that women and older age groups were less likely to report having "felt downhearted". Finally, those with less education (Perkins, et al.), low income, and women (Yu, et al.), were more likely to report having "been a happy person".

Short Form measures – general health items: All five general health items showed DIF for at least one group across two studies. "Health in general" was rated significantly lower by Danes than by Americans (Bjorner, et al., 1998), older people (70+) compared to younger (18-39); those in the low education (0-11 years) as contrasted with those with high education (13+ years), and Blacks vs. Whites (Perkins, et al.). In both data sets examined by Perkins, older people, and those 55-69 years old compared with those 18-39 years old, were less likely to report "getting sick easier". In both data sets, older people were less likely to 'expect their health to get worse'. In the sample from the sick population "expect health to get worse" showed DIF for gender, age and race (Perkins, et al.). Additionally, Perkins, et al. found the item "health is excellent" to show gender DIF in the sick population; males were more likely to endorse the item. DIF was also observed for "I am as healthy as anybody I know" (Bjorner, et al.).

Short Form measures – pain, vitality items: Those aged 70 and older reported less pain than younger people (Fleishman & Lawrence). Perkins and colleagues found that all items in the vitality subscale showed either uniform or non-uniform DIF for age in both data sets examined. Non-uniform findings were more pronounced for those at the lower range of the vitality scale. Conditional on vitality, older people reported having less 'energy' and having felt less "full of pep"; however, they were less likely to report having "felt worn out", and "felt tired". In addition, "felt worn out" showed DIF effects for education and "felt tired" showed non-uniform DIF for race (Perkins, et al.). The examination of the SF-12 showed women were more likely to report not "having energy"; this item was also rated more highly (less symptomatology) by older people, Blacks and Hispanics (Fleishman & Lawrence).

Short Form measures – roles and social functioning items: Physical and emotional role items infrequently exhibited DIF. The physical and emotional role limitation items of the SF-12 showed small effects for education. Those with low education (<12 years) reported that they "accomplished less" (Fleishman & Lawrence). Perkins and colleagues found one physical role item showed positive DIF ("limited in kind of work") for females in the sick population, and older people were more limited in the sick and general populations (Perkins, et al.). An examination of the SF-12 "social activities" item showed older people reported more interference with "social activities" than those younger than 40 and less interference for 'other' race (Fleishman & Lawrence).

Other general health measures: Another general health instrument, the Stanford Health Assessment Questionnaire (HAQ; Fries, Spitz, Kraines & Holman, 1980) was evaluated for Turkish and American populations. The authors found that the item "grip" showed DIF by gender, and the three item Activities subscale showed DIF by culture; Turkish respondents scored slightly higher than did American respondents (Kucukdeveci et al., 2004). Hahn and colleagues (2005) examined uniform DIF for items in the three subscales of the Functional Assessment of Cancer Therapy – Breast (FACT-B; Cella, 1997) for Austrian and American patients using a one-parameter Rasch measurement model, and item location comparison. The trial outcome index (TOI) showed that Americans responded significantly more positively to "enjoy life" and "feel sexually attractive", and significantly less positively for "bothered by weight change", "energy", and "arms swollen and/or tender". In the social/ family well-being subscale (SWB), "family communication" was more negatively rated by Austrians, and "satisfaction (sex life)" was more negatively rated (had a higher threshold calibration) by Americans. Three items from the emotional well-being scale (EWB), "proud of coping", "worry (dying)", and "sad" showed DIF between the groups. Adjusting for DIF did not alter the direction of differences, but slightly altered the effect sizes for each group for all three scales. Teresi and colleagues (2007), examining several FACT items, found that "able to enjoy life" and "content with my quality of life" were more severe indicators for men, and that "worried about dying" was a more severe indicator for the younger cohort and for Blacks.

One additional general health measure, the Sickness Impact Profile (Bergner, Bobbitt, Carter & Gilson, 1981), was examined for DIF. The authors adjusted for sickness level and examined differences between age groups. Younger participants were less likely to endorse "I get around only by using a walker, crutches" and "I do not walk up or down hills". Younger men were less likely to endorse the mobility item "I do not get around in the dark or in unlit places without someone's help" (Lindeboom et al., 2004).

Impact of DIF on general health measures: Half of the studies of health measures examined impact; one (Hahn et al., 2005) examining the FACT-B (Cella, 1997), resulted in the conclusion that impact was slight. The authors (Bjorner, et al., 1998) of a study of the SF-36 concluded that the impact of DIF at the scale level was slight with respect to comparisons of Danish vs. American translations; however, the authors of another study (Fleishman & Lawrence, 2003), examining the SF-12 (Ware, et al., 1996), concluded that the impact of DIF with respect to self-reported Black status and age was large enough to be important. Crane, Gibbons, Narasimhalu, et al. (2007) also found scale level impact related to race for all subscales of the FACT. While several SF items showed significant DIF, removal of items was not recommended by the authors of these studies.

Summary of findings

Review of articles examining DIF in patient self-reported outcomes across three domains (depression, quality of life and general health) identified several poor-performing items within and across domains. These items were flagged because they evidenced large magnitude and/or consistent DIF across studies. Examining first items measuring depression, it was observed that 7 items were problematic. The "crying" item, despite slight semantic divergence across various depression scales (e.g., CESD, Short Care, BDI), showed consistent

DIF for demographic and health-related groups. Additional poor performing items: “people dislike me” (CESD), “everything was an effort” (CESD), “sleep disturbance” (Short Care, DIS, BDI, PHQ-9), “feeling calm and peaceful” (SF) and “appetite” (CESD, BDI, PHQ-9) showed differential functioning for various demographic and/or health-related groups. Similarly, “energy” included in depression (GDS, PHQ-9) and general health (SF) scales showed consistent differential item functioning with regard to several demographic variables.

Turning to items measuring quality of life, worth noting are the items, “felt worried” and “worried about dying”, which are included in emotional distress (FACT) subscales as well as in quality of life (EORTC QLQ-30) measures. A variant of the “worry” item was repeatedly found to perform differently for racial/ethnic, age, and language groups. Also contained in the quality of life measure, EORTC, and in health measures such as the SF and the HAQ are items related to walking. The problematic items are “taking a long walk”, “taking a short walk”, “I find it difficult to walk to the shops”, “walking more than a mile”, “walking several blocks”, and “I do not walk up or down hills”. The general health item “vigorous activities” also showed DIF for a wide variety of demographic variables.

Discussion

Because the following points apply to most of the studies, they are given as an overview prior to review of each content area. Ten of the analyses used a one-parameter Rasch model, six used MIMIC, and three used a M-H or contingency-based method; therefore over half of the analyses were not capable of examining non-uniform DIF. However, a few studies using the Rasch model were able to examine a form of non-uniform DIF by applying the ANOVA approach using Rasch logits or residuals. The use of the ANOVA approach allowed an interaction term of studied group by ability levels, which is an estimate of non-uniform DIF. However, most Rasch analyses were based on t-tests, accompanied by plots of difficulties with 95% confidence intervals. These analyses permitted examination only of uniform DIF. Some of the analyses were very thorough, examining assumptions and impact. Few studies using the Rasch approach examined magnitude of DIF or incorporated purification. A frequently cited reason for the use of the model was the requirement of smaller sample sizes; however, some of the subgroup sample sizes were most likely too small for generalization, e.g., around 30. Generally, the studies that focused on DIF, rather than examining DIF as a byproduct of other analyses produced more comprehensive results.

Few studies employed a two-parameter IRT model that does allow the detection of non-uniform DIF, and most of the authors of studies using the one-parameter model did not test the fit of that model against alternative models with more parameters. No studies used a three parameter IRT model. Unlike educational testing, guessing is rarely examined in studies of DIF in health and psychology. However, as pointed out by Kubinger and Gottschall (2007), guessing behavior may be culturally determined. While it is doubtful that respondents will guess in answering health or mental health questions, it could be a factor in cognitive assessments. Additionally, other response sets, such as a tendency to use the extreme or positive ends of the response category continuums can play a role in biasing item response (McHorney and Fleishman, 2006). It has been found for example that Latino respondents tend to endorse the extreme categories (Marin, Gamba and Marin, 1992). Such factors may help to explain findings related to DIF. Azocar and colleagues (2001) provide a detailed discussion of DIF related to extreme endorsement among Spanish-speakers as contrasted with English speakers.

A handful of studies used the MIMIC latent variable model; the authors applying this approach were meticulous in the execution of the analyses, usually examining assumptions, magnitude and impact. This method allowed simultaneous examination of multiple exogenous (studied) variables. The drawback to the approach is that the method does not allow detection of non-uniform DIF.

One study applied a non-parametric IRT method. A strength of the Mokken scale procedures used by Moorer and colleagues (2001) is that DIF is examined across scales; however, as others have shown, parametric methods of DIF detection might have identified more items with DIF. In that study, no evidence of DIF was observed across disease groups.

The logistic regression approach was also used infrequently; however when applied, it allowed estimation of non-uniform DIF because an interaction term for group and ability could be included. However, the downside of these analyses was that observed scores were typically used as conditioning variables, rather than the theoretically preferred latent variables. While it has generally been advised that purification be used to avoid false DIF detection, this practice was infrequently applied.

Limitations of the analyses

An important limitation is that in the interest of parsimony, the studies reviewed here focused only on formal tests of DIF. However, strict metric-level factorial invariance (using a multi-group factor model) is equivalent to DIF testing using a 2-parameter IRT model (Meredith & Teresi, 2006). Thus, there are measures that have been evaluated using an equivalent method that are not reviewed here. Additionally, although it is believed that the search was comprehensive, it is possible that some studies not contained in the databases searched were inadvertently omitted. A few studies were not included in the table because the DIF analysis was very limited, and could not be evaluated, based on the information presented in the article, or the sample sizes were too small. One study using two methods (Crane, Gibbons, Ocepek-Welikson, et al., 2007; Teresi et al., 2007) was not included in the table because the item set was from several scales; however, the findings were summarized in the body of the paper. Finally, the review was focused only on three areas of patient-reported outcomes, albeit highly salient in terms of clinical trials and observational studies. As previously stated, there were numerous articles on DIF related to function and disease-specific health that were not included in the interest of parsimony.

Summary

Most studies (8 of 14) reviewed in the area of depression examined magnitude, and all but one estimated the impact of DIF; about one-third examined non-uniform DIF. In general, findings were of large amounts of DIF of sizeable magnitude and impact. Adjustments of scale scores were frequently recommended. Among the quality of life studies reviewed, just over half included an assessment of magnitude and of impact of DIF. The authors of these studies tended to conclude that the impact of DIF was minimal, and scale adjustments were not warranted. However, DIF may have been underestimated, as only two studies included a formal evaluation of non-uniform DIF. Finally, of the studies of general health reviewed, half included measures of magnitude; impact was discussed with respect to over half of the studies; however, formal tests of non-uniform DIF were rarely performed. The authors of one of the five studies reviewed that examined the impact of DIF in general health measures concluded that DIF had an important impact on the emotional (mental) health component (Fleishman & Lawrence, 2003). Within the limitations of this review, it might be concluded that depression measures are more subject to DIF than are other types of measures; however, a major caveat is that most of the analyses of general health measures reviewed here did not incorporate tests of non-uniform DIF.

In summary, as a whole, these studies provide good beginning estimates of the presence of DIF in measures of patient-reported outcomes; however, most results should be cross-validated with other (usually larger) samples, using a method that permits examination of non-uniform DIF, while also incorporating the use of latent variables. Examination of magnitude and impact, coupled with qualitative review of item content is also critical in order to achieve an understanding of the role of DIF in assessment of patient-reported outcomes.

References

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleischman, S. B., & de Haes, J. C. (1993). 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, 85,* 365–376.

Albert, S. M. & Teresi, J. A. (2002).Quality of life, definition and measurement. In D.J. Ekerdt (Eds.), *Encyclopedia of Aging Vol. 4* (pp. 1158-1161). NY: Macmillan Reference, Thompson Gale, Inc.

Azocar, F., Areán, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology, 57,* 355–365.

Beck, A. T., Guth, D., Steer, R. A., & Ball, R. (1997). Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for primary care. *Behaviour Research Therapy, 35,* 785-791.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. Psychological Corporation, San Antonio, TX.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care, 19,* 787-805.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51,* 1189-1202.

Bjorner, J. B., Petersen, M. A., Groenvold, M., Aaronson, N., Ahlner-Elmqvist, M., Arraras, J. I., Bredart, A., Frayers, P., Jordhoy, M., Sprangers, M., Watson, M., & Young, T. (2004). Use of item response theory to develop a shortened version of the EORTC QLC-C30 emotional functioning scale. *Quality of Life Research, 13,* 1683-1697.

Bloche, M. G. (2004). Health care disparities – Science, politics, and race. *New England Journal of Medicine, 350,* 1568-1570.

Bock, R. D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 115-122). Hillsdale, NJ: Lawrence Erlbaum, Inc.

Bode, R. K., Lai, J-S., Dineen, K., Heinemann, A. W., Shevrin, D., Von Roenn, J., & Cella, D. (2006). Expansion of a physical function item bank and development of an abbreviated form for clinical research. *Journal of Applied Measurement, 7(1),* 1-15.

Bollen, K. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305-314.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15,* 113-141.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44* (Suppl. 11), S176-S181.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26,* 433-450.

Breslau, J., Javaras, K. N., Blacker, D., Murphy, J. M., & Normand, S. L. T. (2008). Differential item functioning between ethnic groups in the epidemiological assessment of depression. *Journal of Nervous and Mental Disease, 196,* 297-306.

Broekman, B. F. P., Nyunt, S. Z., Niti, M., Jin, A. Z., Ko, S. M., Kumar, R, Fones, C. S. L., & Ng, T. P. (2008). Differential item functioning of the Geriatric Depression Scale in an Asian population. *Journal of Affective Disorders, 108,* 285-290.

Callahan, C. M., & Wolinsky, F. D. (1994). The effect of gender and race on the measurement properties of the CES-D in older adults. *Medical Care, 32,* 341-356.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Castro-Costa, E., Dewey, M., Stewart, R., Banerjee, S., Huppert, F., Mendonca-Lima, C., Bula, C., Reisches, F., Wancata, J., Ritchie, K., Tsolaki, M., Mateos, R., & Prince, M. (2008). Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. *International Journal of Methods in Psychiatric Research, 17,* 12-29.

Cella, D. (1997). *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT Scales)* (Version 4). Evanston, IL: Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare and Northwestern University.

Cella D. (1994). *Manual for the Functional Assessment of Cancer Therapy (FACT) and Functional Assessment of HIV Infection (FAHI) Scales (Version 3).* Chicago: Rush-Presbyterian-St. Luke's Medical Center.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) Scale: An item response theory analysis. *Medical Care, 42,* 281-289.

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

Cohen, P., Cohen, J., Teresi, J., Marchi, P., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14,* 183-196.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *Journal of Clinical Epidemiology, 53,* 285–289.

Copeland, J. R. M., Kelleher, M. J., Kellet, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, J. L., & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule. *Psychological Medicine, 6,* 439-449.

Covic, T., Pallant, J. F., Conaghan, P.G., & Tennant, A. (2007). A longitudinal evaluation of the Center for Epidemiologic Studies scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. *Health and Quality of Life Outcomes, 5,* 41-48.

Crane, P. K. (2006). Commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15,* 1117-1118.

Crane, P. K., Gibbons, L. E., Jolley, L., & Van Belle, G. (2006). DIF analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44* (Suppl. 11), S115-S123.

Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J. S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: the Functional Assessment of Cancer Therapy. *Quality of Life Research, 16,* 101-114.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R.D., & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research, 16,* 69-84.

Crane, P. K., Van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23,* 241-256.

De Jong, Z., Van der Heijde, McKenna, S. P., & Whalley, D. (1997). The reliability and construct validity of the RAQoL: A rheumatoid arthritis-specific quality of life instrument. *British Journal of Rheumatology, 36,* 878-883.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Inc.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: a pragmatic approach. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum, Inc.

Duncan-Jones, P., Grayson, D. A., & Moran, P. A. P. (1986). The utility of latent trait models in psychiatric epidemiology. *Psychological Medicine, 16,* 391-405.

Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6,* 139-150.

Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research, 6,* 393-406.

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning? *Medical Care, 41* (Suppl.), III75-III86.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journals of Gerontology: Social Sciences, 57B,* S275-S284.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14,* 2277-2291.

Foley, K. L., Reed, P. S., Mutran, E. J., & DeVellis, R. F. (2002). Measurement adequacy of the CES-D among a sample of older African-Americans. *Psychiatry Research, 109,* 61-69.

Folstein, M., Folstein, S., & McHugh, P. (1975). Mini-Mental State: a practical guide for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Fries, J. F., Spitz, P., Kraines, R. K., & Holman, H. (1980). Measurement of patient outcomes in arthritis. *Arthritis and Rheumatism, 23,* 137-145.

Gallo, J. J., Cooper-Patrick, L., & Lesikar, S. (1998). Depressive symptoms of Whites and African Americans aged 60 years and older. *Journals of Gerontology, 53B,* P277-P285.

Ganz, P. A., Schag, C. A. C., Lee, J. J., & Sim, M-S. (1992). The CARES: A generic measure of health-related effects of advanced practice nursing on depressive symptoms. *Archives in Internal Medicine, 160,* 2101-2107.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored. An illustration with the Center for Epidemiological Studies Depression Scale. *Educational and Psychological Measurement, 63,* 65-74.

Gibbons, R. D., Clark, D. C., Vonammon-Cavanaugh, S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19,* 43-55.

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire.* London: Oxford University Press.

Golden, R. R., Teresi, J. A., & Gurland, B. J. (1984). Development of indicator scales for the Comprehensive Assessment and Referral Evaluation (CARE) interview schedule. *Journal of Gerontology, 39,* 138-146.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of associations for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

Grayson, D. A., MacKinnon A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiologic Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journals of Gerontology: Psychological Sciences, 55B,* P273–P282.

Gregorich, S.E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(Suppl. 3), S78-S94.

Groenvold, M., Bjorner, J. B., Klee, M. C., & Kreiner, S. (1995). Test for item bias in a quality of life questionnaire. *Journal of Clinical Epidemiology, 48,* 805-816.

Gurland, B.J., Gurland, R.V. (in press). The Choices, Choosing Model of Quality of Life: Linkages to a Science Base. *International Journal of Geriatric Psychiatry.*

Gurland, B. J., Fleiss, J. L., Cooper, J. E., Kendell, R. E., & Simon, R. J. (1969). Cross-national study of diagnosis of mental disorders: Some comparisons of diagnostic criteria from the first investigation. *American Journal of Psychiatry, 125,* 30-39.

Gurland, B., Golden, R., Teresi, J., & Challop, J. (1984). The SHORT-CARE: An efficient instrument for the assessment of depression, dementia, and disability. *Journal of Gerontology, 39,* 166-169.

Gurland, B. J., Yorkston, N. J., Goldberg, K., Fleiss, J. L., Sloane, R. B., & Cristol, A. H. (1972). The Structured and Scaled Interview to Assess Maladjustment (SSIAM): II. Factor analysis, reliability and validity. *Archives of General Psychiatry, 27,* 264-267.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence instrument applicable on adolescents? A latent trait analysis using Rasch-modeling. *Personality and Individual Differences, 36,* 955-968.

Hahn, E. A., Cella, D., Bode, R. K., Gershon, R., & Lai, J. -S. (2006). Item banks and their potential applications to health status assessment in diverse populations. *Medical Care, 44* (Suppl. 11), S189-S197.

Hahn, E. A., Holzner, B., Kemmler, G., Sperner-Unterweger, B., Hudgens, S. A., & Cella, D. (2005). Cross-cultural evaluation of health status using Item Response Theory: FACT-B comparisons between Austrian and US patients with breast cancer. *Evaluation and the Health Professions, 28,* 233-259.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(Suppl. 11), S182-S188.

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-item health survey 1. 0. *Health Economics, 2,* 217-227.

Hepner, K. A., Morales, L. S., Hays, R. D., Edelen, M. O., & Miranda, J. (2008). Evaluating differential item functioning of the PRIME-MD mood module among impoverished Black and White women in primary care. *Women's Health Issues, 18,* 53-61.

Hockwarter, W. A., Harrison, A. W., & Amason, A. C. (1996). Testing a second-order multidimensional model negative affectivity: a cross-validation study. *Educational and Psychological Measurement, 56,* 791-808.

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

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine, 21,* 547-52.

Huelsman, T. J., Nemanick, R. C., & Munz, D. C. (1998). Scales to measure four dimensions of dispositional mood: Positive energy, tiredness, negative activation and relaxation. *Educational and Psychological Measurement, 58,* 804-819.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Jöreskog, K. & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10,* 631-639.

Jorm, A. F., Windsor, T. D., Dear, K. B. G., Anstey, K. J., Christensen, H., & Rodgers, B. (2005). Age group differences in psychological distress: The role of psychosocial risk factors that vary with age. *Psychological Medicine, 35,* 1253-1263.

Katz, S. & Gurland, B.J. (1991). Science of quality of life in elders: Challenge and opportunity. In Birren, J.E., Lubben, J.E., Rowe, J.C., Deutchman, D.E. (eds). *Quality of Life in the Frail Elderly.* San Diego: Academic Press, 335-343.

Kessler, R. C., Wittchen, H -U., Abelson, J. M., Mcgonagle, K., Schwarz, N., Kendler, K. S., Knauper, B., & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the United States. *International Journal of Methods in Psychiatric Research, 7,* 33-55.

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging, 17,* 379-391.

Kind, P. (1996). The EuroQoL instrument: An index of HRQOL. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials.* (2nd ed., pp. 191-201). Philadelphia, PA: Lippincott-Raven.

Koenig, H. G., Meador, K. G., Goli, V., Shelp, F., Cohen, H. J., & Blazer, D. G. (1992). Self-rated depressive symptoms in medical inpatients: Age and racial differences. *International Journal of Psychiatry Medicine, 22,* 11-31.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 169,* 606-13.

Kucukdeveci, A. A., Sahin, H., Ataman, S., Griffiths, B., & Tennant, A. (2004). Issues in cross-cultural validity: Example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. *Arthritis and Rheumatism, 51,* 14-19.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – Some critical suggestions on traditional approaches. *International Journal of Testing*, 5, 377-394.

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on differential item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science, 49*, 361-374.

Kutlay, S., Kucukdeveci, A. A., Gonul, D., & Tennant, A. (2003). Adaptation and validation of the Turkish version of the Rheumatoid Arthritis Quality of Life Scale. *Rheumatology International, 23,* 21-26.

Lai, J. -S., Cella, D., Chang, C. -H., Bode, R. K., Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research, 12,* 485-501.

Lai, J. -S., Dineen, K., Reeve, B. B., Von Roenn, J., Shervin, D., McGuire, M., Bode, R. K., Paice, J., & Cella, D. (2005). An item response theory-based pain item bank can enhance measurement precision. *Journal of Pain and Symptom Management, 30,* 278-288.

Lai, J.-S., Teresi, J., Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and Health Professions, 28, 283-294.*

Lawton, M.P. (1991). A multidimensional view of quality of life in frail elders. In Birren, J.E., Lubben, J.E., Rowe, J.C., Deutchman, D.E. (eds). Quality of Life in the Frail Elderly. San Diego: Academic Press, 3-27.

Lawton, M. P., Kleban, M. H., Rajagopal, D., Dean, J., & Paremelee, P. A. (1992). The factorial generality of brief positive and negative affect measures. *Journal of Gerontology: Psychological Sciences, 47,* P228-P237.

Linacre, J. M. (2005). *Winsteps: Rasch model computer programs.* Chicago, IL: Mesa Press.

Lindeboom, R., Holman, R., Dijkgraaf, M. G. W., Sprangers, M. A., Buskens, E., Diederiks, J. P., & De Haan, R. J. (2004). Scaling the Sickness Impact Profile using item response theory: An exploration of linearity, adaptive use, and patient driven item weights. *Journal of Clinical Epidemiology, 57,* 66-74.

Lopez, A. D., & Murray, C. C. (1998). The global burden of disease: 1990-2020. *American Journal of Public Health, 88,* 196-202.

Lord, F. M., & Novick, M. R. with contributions by Birnbaum, A. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Co.

Makikangas, A., Feldt, T., Kinnunen, Y., Tolvanen, A., Kinnunen, M. L., & Pulkkinen, L. (2006). The factor structure and factorial invariance of the 12-item General Health Questionnaire (GHQ-12) across time: Evidence from two community-based samples. *Psychological Assessment, 18,* 444-451.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

Marin, G., Gamba, R.J. & Marin, B.V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology*, *23*, 498-509.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22,* 357-367.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

McHorney, C. A. & Fleishman, J.A. (2006). Assessing and understanding measurement equivalence in health outcome measures. *Medical Care, 44, Suppl 3*, S205-S210.

McHorney, C. A., Ware, J. E., & Raczek, A. E. (1993). The MOS 36-item short form health survey (SF-36). II. Psychometric and clinical tests of validity in measuring physical and mental constructs. *Medical Care 31,* 247-263.

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,* 361-381.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 302-307.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44* (Suppl. 11)*,* S69-S77.

Millsap, R.E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care, 44* (Suppl. 3), S171-S175.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Moorer, P., Suurmeijer, Th. P. B. M., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research, 10,* 637-645.

Morales, L. S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental Status Exam assessed using the Differential Item and Test Functioning (DFIT) framework. *Medical Care, 44* (Suppl. 11), S143-S151.

Mui, A. C., Burnette, D., & Chen, L. M. (2001). Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS. Measurement in Older Ethnically Diverse Populations. *Journal of Mental Health and Aging, 7,* 137-164.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus User's Guide* (3rd ed.). Los Angeles, CA: Authors.

Nápoles-Springer, A.M., Santoyo-Olsson, J., O'Brien, H., & Stewart, A.L. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44* (Suppl. 3), S21-S30.

National Research Council. (2004). *Measuring racial discrimination.* Panel on methods for assessing discrimination. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

Osborne, R. H., Elsworth, G. R., Sprangers, M. A. G., Oort, F. J., & Hopper, J. L. (2004). 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, 13,* 191-206.

Pagano, I. S., & Gotay, C. C. (2005). Ethnic differential item functioning in the assessment of quality of life in cancer patients. *Health and Quality of Life Outcomes, 3,* 60-69.

Pedersen, R. D., Pallay, A. G., & Rudolph, R. L. (2002). Can improvement in well-being and functioning be distinguished from depression improvement in antidepressant clinical trials? *Quality of Life Research, 11,* 9-17.

Perkins, A. J., Stump, T. E., Monahan, P. O., & McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 Health Survey. *Quality of Life Research, 15,* 331-348.

Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., Fayers, P., Hjermstad, M., Sprangers, M., & Sullivan, M. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research, 12,* 373-385.

Pickard, A. S., Dalal, M. R., & Bushnell, D. M. (2006). 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, 9,* 59-64.

Pratt, L. A., Dey, A. N., & Cohen, A. J. (2007). Characteristics of adults with serious psychological distress, 2001-04. Advance data from *Vital and Health Statistics, 382,* (March 2007). http://www. disc. wisc. edu/reports/cssrindex. html CSSRR #9 (4/3/07).

Prieto, L., Novick, D., Sacristan, J. A., Edgell, E. T., Alonso, J., & SOHO Study Group. (2003). A Rasch model analysis to test the cross-cultural validity of the EuroQoL-5D in the Schizophrenia Outpatient Health Outcomes Study. *Acta Psychiatrica Scandanavica, 107(Suppl. 416),* 24-29.

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401.

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,* 517-528.

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.

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

Reeve, B.B. (2000). Item- and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing Item Response Theory. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 61* (4-B), 2267.

Reeve, B.B. (2006). Special issues for building computerized adaptive tests for measuring patient-reported outcomes: The National Institute of Health's investment in new technology. *Medical Care, 44* (Suppl. 3), S198-S204.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Lui, H., Gershon, R. Reise, S. P., Lai, J.S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 41* (Suppl. 1), S22-S31.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16* (Suppl. 1), 19-31.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38,* 381-389.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64,* 588-599.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* Suppl. 17. Richmond, VA: William Byrd Press.

Schag, A. C., Ganz, P. A., & Heinrich, R. L. (1991). Cancer rehabilitation evaluation system – short form (CARES-SF). *Cancer, 68,* 1406-1413.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomly, A., de Graff, A., Groenvold, M., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research, 16,* 115-129.

Scott, N. W., Fayers, P. M., Bottomly, A., Aaronson, N. K., de Graff, A., Groenvold, M., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2006a). Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15,* 1103-1115.

Scott, N. W., Fayers, P. M., Bottomly, A., Aaronson, N. K., de Graff, A., Groenvold, M., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2006b). Response to commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15,* 1119-1120.

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

Shelvin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17,* 231-236.

Simon, G. E., & Von Korff, M. (2006). Medical co-morbidity and validity of DSM-IV depression criteria. *Psychological Medicine, 36,* 27-36. Epub 2005 Oct 5.

Skinner, J. H., Teresi, J. A., Holmes, D., Stahl, S. M., Stewart, S. L., (Eds.). (2001). Measurement in older ethnically diverse populations [Special issue]. *Journal of Mental Health and Aging, 7*(1).

Smedley, B. D., Stith, A. Y., & Neslon, A. R., (Eds.). (2003). *Unequal treatment: confronting racial and ethnic disparities in health care.* Washington, D. C.: National Academies Press.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *Medical Research Methodology, 8,* 33.

Spitzer, R. L., Fleiss, J. L., Burdock, E. I., & Hardesty, A. S. (1964). The Mental Status Schedule: Rationale, reliability and validity. *Comprehensive Psychiatry, 5,* 384-395.

Spitzer, R. L., Williams, J. B., Kroenke, K., Linzer, M., deGruy, F. V. r., Hahn, S. R., et al. (1994). Utility of the new procedure for diagnosing mental disorders in primary care: The PRIME-MD Study. *Journal of the American Medical Association, 272,* 1749-1756.

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.

Steinbrook, R. (2004). Disparities in health care – From politics to policy. *New England Journal of Medicine, 350,* 1486-1488.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Takane,Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.

Tang, W. K., Wong, E., Chiu, H. F. K., Lum, C. M., & Ungvari, G. S. (2005). The Geriatric Depression Scale should be shortened: Results of Rasch analysis. *International Journal of Geriatric Psychiatry, 20,* 783-789.

Teresi, J. A. (2006a). Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Medical Care, 44* (Suppl. 11), S152-S170.

Teresi, J. A. (2006b). Overview of quantitative measurement methods: Equivalence, invariance and differential item functioning in health applications. *Medical Care, 44* (Suppl. 11), S39-S49.

Teresi, J., Abrams, R., & Holmes, D. (2000). Measurement of depression and depression recognition among individuals with cognitive impairment. In S. Albert & R. Logsdon (Eds.), *Assessing quality of life in Alzheimer's disease* (pp. 121-151). New York: Springer.

Teresi, J., Abrams, R., Holmes, D. Ramirez, M., Shapiro, C., & Eimicke, J. P. (2002). Influence of cognitive impairment, illness, gender, and African-American status on psychiatric ratings and staff recognition of depression. *American Journal Geriatric Psychiatry, 10,* 506-514.

Teresi, J., Cross, P., & Golden, R. (1989). Some applications of latent trait analysis to the measurement of ADL. *Journals of Gerontology: Social Sciences, 44,* S196-S204.

Teresi, J., & Golden, R. (1994). Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. *Alzheimer Disease and Associated Disorders, 8 (Suppl 1),* S291-S298.

Teresi, J., & Holmes, D. (1994). Overview of methodological issues in gerontological and geriatric measurement. In: P. Lawton & J. Teresi (Eds.), *Annual Review of Gerontology and Geriatrics: Focus on assessment techniques* (pp. 1-22). New York: Springer.

Teresi, J. A., & Holmes, D. (2001). Some methodological guidelines for cross-cultural comparisons. *Journal of Mental Health and Aging, 7,* 13-19.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19,* 1651-1683.

Teresi, J., Ocepek-Welikson, K., Kleinman, M., Cook, K.F., Crane, P.K., Gibbons, L.E., Morales, L. S., Orlando-Edelin, M., & Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research, 16,* 43-68.

Teresi, J. A., Stewart, A. L., Morales, L., & Stahl, S. (2006). Measurement in a multi-ethnic society: Overview to the special issue. *Medical Care, 44* (Suppl. 11), S3-S4.

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: Lawrence Erlbaum, Inc.

Ware, J. E., Gandek, B., & The IQOLA Project Group. (1994). The SF-36 Health Survey: Development and use in mental health and the IQOLA Project. *International Journal of Mental Health,* 23, 49-73.

Ware, J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlof, C. G., Tepper, S., & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12,* 935-952.

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item, short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34,* 220-233.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30,* 473-83.

Wang, W.C., Yao, G., Tsai, Y. J., Wang, J. D., & Hseih, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research, 15,* 607-620.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59,* 910-927.

Wright, B. D., & Stone, M. (1979). *Best Test Design.* Chicago IL: Mesa Press.

Yang, F. M., & Jones, R. N. (2007). Center for Epidemiologic Studies – Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology, 60,* 1195-1200.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of Psychiatric Research, 17,* 37-49.

Yu, Y. F., Yu, A. P., & Ahn, J. (2007). Investigating differential item functioning by chronic diseases in the SF-36 Health Survey. *Medical Care, 45,* 851-859.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67,* 361–370.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www. educ. ubc. ca/faculty/zumbo/DIF/index. html.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science. Prince George Canada: University of Northern British Columbia.

Acknowledgements

These analyses were supported in part by the United States National Institute on Aging: Resource Centers on Minority Aging Research (5P30AG015294), the United States National Institutes of Health Roadmap project: Patient Reported Outcomes Information System (PROMIS) (5U01AR052177), and the United States National Institute on Aging project: Understanding Disparities in Mental Status Assessment (AG0253008).

Table 1:
Summary of DIF results for measures of depression, quality-of-life and general health.

Depression

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
Beck Depression Inventory (BDI; Beck et al., 1961) *DIF Method:* Mantel-Haenszel for Ordered Response Categories (Mantel, 1963). An extended standardization procedure using the *z* statistic (Dorans and Schmitt, 1993). The Bonferroni control for multiple comparisons to determine significance was used (Azocar, et al., 2001).	*Non-Uniform:* Not estimated with this method *Uniform:* Four BDI items showed DIF for the comparison of Latino and Anglo respondents. Controlling for level of depression, the items: "I feel like I am being punished" (*z*=2.86, *p* < .003); "I feel like crying" (*z*=2.07, *p*< .05); and "I believe I look ugly" (*z*=2.16, *p*< .01) were more likely to be endorsed by Spanish than by English speakers. On the other hand, the Spanish speakers were less likely to endorse the item "I can't do any work at all" (*z*= -2.34, *p*< .01).	*Magnitude:* Not discussed *Impact:* According to the authors, given equivalent depression scores, Latino samples could have mean scores up to six points greater than English speaking samples. "If a Latino patient is not depressed, yet endorses "I feel I'm being punished", "I feel like crying", and "I believe I look ugly", the BDI score could be as much as nine points higher than a non depressed Anglo patient. Biased items artificially increase or, as with "I can't do any work at all", decrease the total score of the scale." pg 363	*Strengths:* The article identifies two potentially semantically different items in two independent translations of the BDI, and provides a good contextual discussion, anchored in Latino cultural norms, of putative causes of DIF. *Possible Limitations:* 1. The sample size of the focal group (Spanish speaking) was small (n=55). 2. Assumptions of the model were not explicitly discussed. While non-parametric models have few assumptions, the lack of unidimensionality suggested by the factor analyses may be a problem because lack of unidimensionality can result in inaccurate DIF detection. 3. Purification was not performed. 4. Only uniform DIF was tested. 5. The magnitude of DIF was not discussed.
Beck Depression Inventory (BDI; Beck et al., 1961) *DIF Method:*	*Non-Uniform:* Non-uniform DIF was most common (8 of 11 items). Midlife patients endorsed higher scores at low to average levels of depression, with fewer differences at high	*Magnitude:* "The loss of libido item showed the largest DIF (DIF value= 3.13; percentage of the total differential test functioning (DTF) = 12.5%), and the	*Strengths:* 1. According to the authors, the analyses provide additional perspective on the newer revisions of the BDI, such as the BDI-II (Beck, Steer, & Brown, 1996) and the BDI-

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
IRT Log-likelihood ratio test using the 2p graded IRT model. The first parameter (α_i) describes how well an item discriminates between individuals who are lower and those who are higher on the depression dimension. Differences in this parameter indicate non-uniform DIF. "The second (multiple threshold) parameter can be interpreted as an inflection or cutting point between two adjacent response levels." (Pg 380) Differences in this parameter indicate uniform DIF (Kim, et al., 2002).	levels for the items: self-criticism, social withdrawal, irritability, guilt feelings, and sense of failure. The other pattern, in which both groups endorsed similar scores at low levels of depression, with late-life patients endorsing higher scores at more severe levels of depression, was observed for sleep disturbance, somatic preoccupation, and work inhibition. Late-life patients reported fewer cognitive symptoms (disappointment in self, self-criticism, guilt, and sense of failure) at lower levels of depression, and more somatic symptoms (sleep disturbance, somatic preoccupation, weight loss), at higher levels. *Uniform:* According to the authors, "The ICCs for midlife and late-life groups on the mood item (one of four anchor items) overlap, indicating that the item functioned in the same way in both patient groups." (pg 383) Three items with uniform DIF were: "weight loss", "loss of libido" and "disappointment in self". For the latter two items, midlife patients endorsed consistently higher scores than late-life patients. "With the third item (weight loss), late-life patients endorsed higher scores, but this was a poor item in general, with low levels of	ICCs indicate that the item was consistently difficult for late-life patients to endorse regardless of the severity of their depression". The ICCs for the self-accusation (DIF value= 1.94; percentage of the DTF = 7.7%) and sleep disturbance (DIF value= 2.27; percentage of the DTF = 9.0%) items illustrate non-uniform DIF. These items showed the second and fourth largest DIF among the items in the BDI. *Impact:* Approximately half of the items (11 of 21) on the BDI accounted for about 80% of the DTF. "Eleven of the items (loss of libido; sleep disturbance; weight loss; self-accusation; self-dislike; social withdrawal; somatic preoccupation; irritability; work inhibition; guilt feelings; and sense of failure) in the BDI each accounted for 5% or more of the DTF. Six items (sense of punishment; crying; distorted body image; indecisiveness; pessimism; and suicidal wishes) accounted for 1%–5% of the DTF." Cutoff scores adjusted on the basis of the IRT model had an impact on assignment to depression levels.	PC for primary care settings (Beck, Guth, Steer, & Ball, 1997). 2. The authors provide an excellent review, explication, and execution of the likelihood ratio approach to DIF detection. 3. The differences documented in the present analyses show that age is a factor that may contribute to variability in test scores. The possibility that some of the discrepancies in the research literature concerning the prevalence of depression in older versus younger cohorts may be the result of differential functioning of the BDI for late-life patients is highlighted. 4. Sample size was adequate for the exercise. 5. Both uniform and non-uniform DIF were examined. 6. Purification was performed. 7. Possible reasons for age-related differences in the performance of the BDI are discussed. 8. The use of cutoff scores for the late-life group as a way for adjusting for DIF in the BDI was discussed. Cutoff scores were provided. The obvious caveat (as discussed by the authors) is that such adjustments are sample dependent and may not be cross-validated. *Possible Limitations:* There are no obvious

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	**Magnitude & Impact**	
	endorsement (low elevation) and a flat slope (poor discriminating power across levels of depression).” (Pg 384) The somatic anchor items, loss of appetite and fatigability, functioned similarly in both groups, and loss of libido was more characteristic of midlife patients. (Pg 384)	“Using the adjusted cutoffs for late-life patients lowered the false negative rate by more than half (9.6% to 4.6% of clinically diagnosed patients whose scores on the BDI would have placed them in the nondepressed range). The percentage of severely depressed late-life patients also decreased (17.0% to 13.3%), and the percentage of moderately depressed patients increased (41.3% to 49.5%).” (pg 385)	limitations to these analyses, except that this method does not permit inclusion of covariates.
Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) *DIF Method:* Item Response Theory Evaluation of Item-Level Mode Effect: Likelihood ratio tests were used to test whether item parameters were significantly different by interview mode. The two parameter graded response model was applied using	*Non-Uniform:* One item evidenced non-uniform DIF: “felt hopeful about the future”. *Uniform:* Examination of category response curves showed that phone respondents had a lower probability of endorsing the third of four response categories, than did mail respondents. “A higher level of underlying depression is needed for phone respondents to endorse the higher categories.” Twelve items showed uniform DIF. They were: “I was happy”, “I enjoyed life” “I could not get going”, “I talked less than usual”, “I felt everything was an effort”, “I felt that I was just as good as other people”, “I felt depressed”, “I felt sad”, “I had trouble keeping my mind on what I was	*Magnitude:* Not provided *Impact:* Evaluation of test characteristic curves indicated that the mode effect is stronger for persons in the middle range of the depression continuum, with scores for mail respondents up to 6 points higher. “For example, at -1.0 and 0.5 on the Depression scale, the expected scores for mail respondents are approximately 3 and 5 points higher, respectively, than those for phone respondents.” (pg 286) According to the authors, the study suggests that “cognitive processing could be an important contributor to the interview mode effect observed in the CES-D.”	*Strengths:* 1. The use of IRT, and a randomized design to examine the comparability of self- versus interviewer-administered versions of the CES-D is an important contribution. 2. Design features (clustering) were considered and dismissed as requiring modeling. The parallel group design permitted the assumption of equality of population means and S.D.s in the two groups. 3. The impact of DIF was discussed. 4. Score adjustment of the CES-D was discussed in the context of comparisons made across data sources or studies using different modes of data collection. *Possible Limitations:* 1. According to the authors, sample sizes were too small to permit inclusion of

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
MULTILOG (Chan et al., 2004).	doing", "I felt people disliked me", "I thought my life had been a failure", and "I felt hopeful about the future".		interaction terms with social and demographic variables. 2. Subjects who refused to participate, and those who were excluded from random selection showed significantly different demographic characteristics as compared to participants. 3. It was unclear whether or not a purification procedure was performed. 4. DIF magnitude was not discussed.
Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) *DIF Method:* An extension of the Mantel-Haenszel method, a proportional odds regression model was used for polytomous (ordinal) items. The Bonferroni correction was used; the adjusted p-value was <0.0008 (Cole, et al., 2000).	*Non-Uniform:* Non-uniform DIF was not discussed, although it was estimated through the use of the interaction term for CES-D by studied exogenous factor. Neither of the differences in item difficulty varied by level of depressive symptoms (p for interaction=0.53 for both tests). *Uniform:* Three of 20 CES-D items were found to show DIF by age, gender, and racial group. "The mean score for the 'people are unfriendly' item, adjusted for overall depressive symptoms, was 0.37 (standard error (s.e.) 0.02) for Blacks and 0.19 (s.e. 0.01) for Whites. The mean score for the 'people dislike me' item, adjusted for overall depressive symptoms, was 0.25 (s.e. 0.02) for Blacks and 0.11 (s.e. 0.01) for Whites. The mean score for the 'crying	*Magnitude:* An effect size cut-point (odds ratio (OR)>2.0 or <.5) to define meaningful item bias was used. Three items with larger magnitude item-level DIF (OR >2.0 and Spearman rank correlation >0.10) were identified. The odds of Blacks responding higher on the items "people are unfriendly" and "people dislike me" were respectively, 2.29 and 2.96 times that of Whites matched on overall depression. The odds of women responding higher on the item "crying spells" were 2.14 times (95% confidence interval (CI): 1.60, 2.82) that of men matched on overall depressive symptoms. *Impact:* A relatively "DIF-free" 17-item version correlated 0.99 with the full version.	*Strengths:* 1. The authors note that the use of a proportional odds regression model allowed for examination of DIF in item difficulties and discrimination parameters; and 2. A test for differential factor function was performed by examining the relationship of the factor score with each studied (exogenous) variable after conditioning on the CES-D score. 3. This analysis was well executed using a model that can detect both uniform and non-uniform DIF. 4. The authors considered carefully the model assumptions. 5. The sample size was adequate, and effects of purification were examined. 6. Interpretation of the impact of DIF was offered; it was posited that the two IP items might contribute to the association between

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
	spells' item, adjusted for overall depressive symptoms, was 0.23 (s.e. 0.01) for women and 0.15 (s.e. 0.02) for men. This difference in item difficulty by gender did not appear to vary by level of depressive symptoms (p for interaction= 0.86). There was no evidence of any item bias by age group in this sample of elders." (see pg 287)	Using a "standard cutpoint of > 16 points on the full 20-item scale as the threshold for diagnosis of depression, the sensitivity and specificity of the reduced scale 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.	"perception of racial prejudice" and depression. *Possible Limitations:* 1. The authors acknowledge that the lack of evidence of item bias by age might be due to the restricted age range in the sample. 2. There were no formal tests of the assumption of unidimensionality. 3. There was no discussion of types of DIF. 4. An observed rather than latent conditioning variable was used.
Center for Epidemiologic Studies-Depression scale (CES-D; Radloff, 1977) *DIF Method:* The polytomous data were analyzed using a partial credit rating scale model. Thresholds (0.5 probability point between adjacent categories may vary across items. Graphs were used to compare item locations across sub-groups of	*Non-Uniform:* Non-uniform DIF is suspected when significant interaction effects were identified in ANOVA analysis. Findings were not discussed. *Uniform:* Uniform DIF is defined as significant MAIN effect in ANOVA of residuals. The analyses were focused on 13 items from the CES-D, administered to patients with rheumatoid arthritis. Seven items were found not to fit the model and were excluded. These items were "feeling as good as others", "feeling hopeful", "feeling happy", "enjoying life", "feeling sad", "poor appetite", and "restless sleep". Using Bonferroni-adjusted p value,	*Magnitude:* Not provided *Impact:* Not provided	*Strengths* 1. Both qualitative (i.e., graphic) and quantitative (i.e., statistical comparison) were used. *Possible Limitations:* 1. The authors did not explain explicitly their method and rationale. 2. Although the authors mentioned the "graphic approach" and the ANOVA DIF, details were not provided and tests of non-uniform DIF were either not performed or not discussed. 3. DIF analysis assumes unidimensionality. This should be tested prior to examining DIF. 4. The sample size was small, particularly for subgroup analysis of polytomous items, and the findings may not be robust.

Measure/ DIF Method	**DIF Results**		**Review**
	Non-uniform & Uniform	**Magnitude & Impact**	
interests (i.e., time-points, gender and age). ANOVA embedded in RUMM 2020 was used to examine DIF (Covic et al., 2007).	significant uniform DIF for both age & gender were identified. "I felt tearful" and "I had crying spells" had significant DIF for both age and gender. The sample aged 53 yrs or less (compared to 54-65 & 66+) and female (compared to male) were significantly more likely to endorse these two items.		5. The Rasch model may not have been the most appropriate model for the data. 6. Purification was not performed. 7. Magnitude and impact of DIF was not examined. 8. Although the article discussed "tearful" throughout, the CES-D item is "fearful". It is unclear whether the actual item used in the study was "tearful" or "fearful".
Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) *DIF Method:* The Multiple Indicators, Multiple Causes (MIMIC) Model (Muthén, 1984) The measurement model specified that the 20 CES-D items load on a single latent depression factor. The model included multiple regression of the latent variable on covariates by specifying disability,	*Non-Uniform:* Cannot be tested with the MIMIC model. *Uniform:* Education, divorced, and obesity had critical ratios that did not exceed either 1.5 on the predictor loading to the CES-D factor or 2.0 for any of the 19 bias loadings. Item-specific effects were found: "Older participants reported being more bothered by things and less hopeful about the future; men found things less of an effort, were less fearful, slept better and reported crying less. Being widowed was associated with feeling at least as good as others and with more fear and loneliness." Mobility, ADL, and IADL also had direct effects on poor appetite, finding everything an effort, restless sleep, and inability to get going. "The items 'good as others,' 'talked less,' 'people unfriendly,' 'enjoyed life,' 'crying spells,' 'felt sad,' and 'people dislike me' all showed significant	*Magnitude:* Item-level magnitude could be estimated by examination of the β_i or the path from the studied covariate to the item. "Bias effects of age on CES-D of 0.06 arise from the only significant loadings, which are on the items 'bothered by things' (0.02) and 'hopeful about the future' (0.04); whereas age has no loading on the Depression factor." None of the demographic effects were associated with elevated depression. Of the disability variables, the increases on the depression factor were associated with any disability other than ADL. "Physical disorder variables also influenced the CES-D Depression factor and items directly. Heart disease, stroke, any other systemic disease, gait instability, and cognitive impairment	*Strengths:* 1. This is an innovative parametric method that uses latent variable models to examine the DIF effects. The analyses were well-executed. 2. Both magnitude and impact of DIF were examined. *Possible Limitations:* Limitations noted by the authors relate to the choice for identification purposes of the item 'felt depressed' as unbiased. "The factor of depression then becomes that for which this item is unbiased, and biases on other items are in relation to this particular factor. However, had 'everything an effort' been selected to be unbiased, the corresponding factor of depression would have absorbed these effects, showing positive associations between these physical disorders and the (new) factor of depression, whereas the symptom 'felt depressed' would now show

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
physical disorder and demographic characteristics as predictors of latent depression. “ In addition to estimating the effect of the covariates on the latent variable (γ), the model allows the effects of the covariates on each of the CES-D items to be assessed directly.” (pg 274) The reference indicator, “felt depressed”, was constrained for identification to have zero bias (Grayson, et al., 2000).	negative direct loadings; individuals with physical disorders underreport on these items for reasons unassociated with depression. As with disability, the items ‘poor appetite,’ ‘everything an effort,’ and ‘inability to get going’ had higher endorsement levels in individuals with particular diseases for reasons other than a disease-related elevation in depression.” (pg 276). “Individuals with a disability, bone and joint disease, and stroke were more likely to report that ‘everything is an effort’ above and beyond levels of depression. The items ‘good as others,’ ‘talked less,’ ‘people unfriendly,’ ‘enjoyed life,’ ‘crying spells,’ ‘felt sad,’ and ‘people dislike me’ had negative associations with particular disorders; participants with more severe physical disorder respond to these items in a less extreme manner than expected for given depression levels.” (pg 279)	were all associated with a genuine rise in depression, although they all show other effects on the CES-D” (pg P276). *Impact:* The bias effect (impact) of the items on the total score is estimated by the sum of all the direct loadings (Σ β_i) from the studied variable to the 20 items. The genuine effect of the predictors on the latent variable is estimated as the sum of the loadings from the depression factor to the items (Σ λ_i) multiplied by the path from the predictors to the depression factor (γ). “The regression of age on CES-D total score yields a beta weight of 0.09: Each increase of 1 year in age is associated with an increase of 0.09 in CES-D total.” “The bias effects range from negligible (widowed) to over seven times the magnitude (bone and joint disease) of the effects on the Depression factor (averaging 157%) and are frequently in the opposite direction. Using tests of significance: Of 17 predictors, 9 show only nondepression effects on the CES-D, and only 1 (incontinence) supports the use of the CES-D as an unbiased measure of depression; with the remaining 7	biases indicating underreporting.” (pg P279, P281). (It is noted by the reviewer that while the assumption that the item constrained to identify the scale (reference indicator) is DIF-free might not be valid; the necessity and the effects of “purification” of the item selected for identification is an area of controversy, and requires further study.) *Additional possible limitations are*: 1. Sample sizes for the different variables were not reported; however, based on the overall n, the sample size might have not been adequate. 2. Multicollinearity among covariates was not discussed. 3. The main limitation is the inability to examine non-uniform DIF.

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
		predictors showing joint contributions, the bias component ranges from 4% (heart disease) to 64% (mobility) of the magnitude of the genuine depression component." (pg 278)	
Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) *DIF Method:* The 1-parameter Rasch model (Rasch, 1960; Pickard et al., 2006)	*Non-Uniform:* Cannot be tested with this method. *Uniform:* Four items demonstrated statistically significant DIF: "my sleep was restless," "I felt that people disliked me," "I did not feel like eating," and "I had crying spells." The authors conclude that DIF observed between depressed stroke and primary-care patients may imply that slightly different clusters of depressive symptoms occur in stroke compared with primary-care patients. They note that the same items have been associated with bias in studies of people without stroke.	*Magnitude:* A logit difference of approximately 0.5 or more across the two groups was used as a measure of magnitude. The logit differences ranged from .77 to .03. Items with the largest differences were: disliked (.77), appetite (.65), restless (.61), crying (.48), as good as others (.48). *Impact:* Only one item (not specified) was identified as uniquely, psychometrically problematic in the stroke subgroup; the authors do not recommend stroke-specific changes to the CES-D scale.	*Strengths:* The authors provide a discussion of the DIF results, anchored in findings from other studies. *Possible Limitations:* 1. The authors recognized limitations in generalizability due to sampling issues. 2. The timing of the assessment, (three months poststroke) may be problematic. 3. The small size (n=32) of the stroke group may have led to low power to detect DIF. 4. Non-uniform DIF was not examined. 5. Relevant covariates could not be modeled. 6. Tests of unidimensionality were not performed, and the authors concede that the factor structure of the CES-D in nonstroke patients has been found to have four underlying factors (negative and positive affect, somatic and interpersonal disruption.)
The Composite International Diagnostic Interview (CIDI; Kessler et al., 1998)	*Non-Uniform:* NA *Uniform:* Whites, Blacks & Hispanics were compared using a US general population sample.	*Magnitude:* Not discussed *Impact:* Results from analyses inclusive of items with DIF were compared by using chi-	*Strengths:* 1) Hypothesis-driven analyses were used to determine benign and adverse DIF at both item and symptom levels. 2) Though DIF was explained/illustrated by

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
DIF Method Nonparametric IRT using SIBTEST. DIF is conceptualized as a difference in the probability of endorsing an item response, occurring when individuals in groups having the same levels of the latent construct, possess different amounts of nuisance abilities that influence responding. If the p <0.05 from a 1-sided test, the item was flagged as exhibiting DIF (Breslau et al., 2008).	*White and Blacks* (bias toward underestimation of depression among Blacks relative to Whites) *At item level*: "lack of energy" (p=.002), "felt worthless' (p=.002), "thoughts of suicide" (p<.001) *At symptom level*: "loss of energy" (p=.002) & "self-reproach" (p<.001) *White and Hispanics* (underestimation of depression among Hispanics) *At item level:* "increased weight" (p=.003) & "waking early" (p=.001) *At symptom level*: "suicidality" (p=.003)	square tests. Specifically, 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 depressive episode among Whites and Blacks was: (with all items) 18.6% and 12.6% (χ^2=10.69, p=.002); (after the adjustment for DIF) 17.7% and 12.4% (χ^2=8.20, p=.007), respectively. The prevalence estimates for Whites and Hispanics were similar, about 18% both before and after item removal.	using 2-P IRT based ICCs, a non-parametric approach was chosen with the assumption of no known underlying distribution. This is an acceptable rationale. 3) Purification was conducted by examining DIF before and after removal of items with DIF. *Possible Limitations:* 1) It would have been of use to discuss the DIF magnitude. 2) SIBTEST does not permit detection of non-uniform DIF. 3) Use of a one-sided DIF procedure in purification is counter to recommended practice. The authors argue that use of a one-sided test of DIF during purification protects against unnecessary removal of DIF-free items; however, the procedure could also result in a less pure anchor.
Major Depression in the Diagnostic Interview Schedule (DIS; Robins, Helzer, Croughan & Ratcliff, 1981) *DIF Method:* The MIMIC model (a	*Non-Uniform:* MIMIC models do not permit examination of non-uniform DIF. *Uniform:* At the Baltimore site, the estimates for the direct effect of ethnicity on depression items indicate several items with DIF. Sleep disturbance was less likely to be endorsed by African Americans (AAs) than Whites,	*Magnitude:* The magnitude of the direct effects are presented in Table 5, but are not discussed. *Impact:* The one month prevalence estimates for symptoms were higher among AAs than among Whites in Baltimore, but tended to be lower among AAs compared with	*Strengths:* 1. The authors noted several advantages of MIMIC for DIF detection by self-reported race. 2. The discussion section is an excellent presentation of other related findings and of the possible differences in meaning and cultural world view that may affect findings.

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
special application of the latent trait model) (Muthén & Muthén, 1998-2004) was used for DIF detection. Covariates to the MIMIC model were introduced in order to examine DIF by self-reported race, adjusting for differences in the level of the latent trait (depression) and for the effect of other covariates such as gender, educational status, MMSE score (Folstein, Folstein & McHugh), and marital status (Gallo et al., 1998).	given level of depression and the other covariates. Difficulty concentrating and thoughts of death were more likely to be reported by older AAs one month prior to interview, when compared with older Whites. Older AAs were less likely to report other symptoms, but the differences were not significant. “Sadness” was less likely to be endorsed by AAs at both sites. Three symptoms demonstrated statistically significant differences in AAs compared with Whites at the Durham-Piedmont site; sadness, loss of interest, and thoughts of death. While older AAs were more likely to endorse the item bundle of “thinking about death or suicide” than were Whites, the main difference in endorsement arose from the item about thoughts of death. (pg 280-282)	Whites in Durham-Piedmont. According to the authors, “significantly higher regression estimates (of depression on the covariates of race, gender, education, MMSE score, and marital status) for AAs at the Baltimore site infer that older AAs are higher on the latent trait of depression when compared with older Whites.” (pg 280) At the Durham-Piedmont site, mean levels of depression were lower for older AAs than for Whites. Women, those with lower education, and unmarried persons had lower levels of depression.	*Possible Limitations:* 1. The authors note that the validity and assessment of the concept of race can be questioned. “While differences ascribed to ‘race’ may reflect social more than genetic differences, failure to account for heterogeneity in measurement of depressive symptoms might be misleading in comparing prevalence across ethnic groups.” The extent to which African Americans ascribed their symptoms to a physical illness is unknown because chronic illnesses were not modeled. (pg 283) 2. The authors note that racial or ethnic background of the interviewer and sampling fluctuations may have affected the results. 3. Inability to model non-uniform DIF by examining differences in the slope parameter is a limitation. 4. More discussion of the model assumption and DIF magnitude tests would have been useful.
The Geriatric Depression Scale (GDS) *DIF Method*: Multiple Indicator, Multiple Cause model (MIMIC). Comparisons	*Non-Uniform:* Cannot be tested with the MIMIC model *Uniform:* DIF was examined in this sample of chronic elderly in Singapore. Six items showed age-related DIF: The older old were more likely to endorse ‘drop many activities and interests’, ‘feel (not) happy’, ‘prefer staying	*Magnitude:* Ten items evidenced DIF on one or more factors, two of which had a MI of borderline magnitude for gender or age. After deleting those two, eight (of the 15) items with significant DIF remained. These were: ‘dropped many activities and interests’, ‘afraid	*Strengths:* 1. The study sample, a large multi-ethnic (Chinese, Malay and Indian) Asian, community-residing older persons represents a contribution to the field. 2. DIF identification with respect to age, gender, ethnicity and chronic illness is instructive when performing comparisons of

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
of baseline no-DIF and DIF models represented by changes in the Modification Index (MI) statistic were performed. MI is related to misfit, and tested as a Chi-square statistic with one df. (Broekman, et al., 2008).	home', 'more problems with memory', 'feel pretty worthless' and 'not full of energy'. Five items showed gender-related DIF: '(not) satisfied with life', 'afraid that something bad is going to happen', 'prefer staying home', 'more problems with memory' and 'feel situation is hopeless' in the direction of females being more likely to report symptoms. Males were more likely to report "dissatisfaction with life". Four items showed ethnicity-related DIF: 'prefer staying home', 'think not wonderful to be alive', 'feel pretty worthless' and 'more problems with memory'; with Chinese respondents more likely to report symptoms. Two items showed illness-related DIF: 'feel pretty worthless' and 'not full of energy'; those with chronic illness were more likely to report symptoms.	something bad is going to happen', 'prefer staying home to going out', 'more problems with memory than most', 'think it is (not) wonderful to be alive', 'feel pretty worthless', 'feel (not) full of energy', and 'feel that situation is hopeless'. *Impact:* The effects of DIF related to age, gender, ethnicity and chronic illness could potentially bias the GDS-15 scores. Age-related DIF appeared to have the greatest and chronic illness the least DIF impact at the scale level. DIF cancellation (DIF occurring in opposite directions for different items) was observed. This resulted in the masking or cancellation of bias at the scale level (see pg 288).	prevalence of depressive symptoms using GDS scores across Asian population subgroups. *Possible Limitations:* 1. The authors point out the smaller number of minority Indian and Malay subjects which had to be combined for analysis, and the self-report of chronic medical illnesses as study limitations. 2. The inability to examine non-uniform DIF is a methodological limitation of the study.
The Geriatric Depression Scale (GDS; Yesavage et al., 1982) *DIF Method:* Unidimensionality, item fit, and DIF were assessed using the Rasch models in WINSTEPS, Version	*Non-Uniform:* The authors defined non-uniform DIF as the difference between the groups in the probability of a positive item response that varies across the trait (ANOVA interaction) Because there was no significant interaction in the ANOVA of residuals, the authors conclude that non-uniform DIF was not detected. *Uniform:* For each item, standardized residuals of the	*Magnitude:* Not discussed *Impact:* The four items exhibiting DIF were deleted in order to create a new version of the scale. There was no significant difference between the area under the ROC curves (AUC) of the original and revised versions of the GDS.	*Strengths:* Use of ANOVA with residuals provides additional information about DIF, beyond that provided by fit statistics and t-tests of the difficulty parameters. *Possible Limitations*: 1. The authors recognized the limited generalizability of the results due to the study sample ("findings may not be applicable to non-Chinese populations or patients with other forms of medical

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
3.04 (Linacre, 2005), with an unconditional maximum likelihood procedure. Bonferroni corrections were used to adjust for multiple comparisons (Tang et al., 2005).	observed from model predicted scores are calculated. The test of DIF is an ANOVA of the person-item deviation residuals with person factors (e.g., age & education) and class intervals (e.g., group along the trait) as factors. Uniform DIF was defined by the authors as a "constant difference between groups in the probability of affirming an item (or category) across the trait (ANOVA main effect)." "Person and item separation index were 1.80 and 4.54, respectively. No items had a significant DIF for age, education and cognitive impairment."		disease, such as stroke"). 2. Additionally, the DIF analyses included only small numbers of individuals with low MMSE's (below 15), and should be replicated. 3. Sample sizes for the group variables (age, education, cognitive impairment) were small. 4. Purification was not performed. 5. Model fit was not compared to a 2-parameter IRT model; the Rasch model may not have been the best choice if discrimination parameters differ across groups.
Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983) *DIF Method:* Restricted factor analyses using Structural Equation Modeling (SEM). "Initial exploration of item function was conducted using descriptive analysis: Mann Whitney U test." DIF was defined as a	*Non-Uniform:* Not provided *Uniform:* *Anxiety:* The baseline model was expanded by the addition of a measured independent variable for the difference between population reference women and those diagnosed with breast cancer. Model modification tests were examined for adding correlations between this group variable and the residuals of the anxiety items. The final and the baseline models, which included the group variable, were 'nested'; therefore, the change in goodness-of-fit could be tested directly and was found to be significant (incremental X^2 =39.2, 2	*Magnitude:* The authors report low magnitude of DIF: the largest direct correlation between an item residual and the group variable was -.20, representing approximately 4% of the variance of the item": 'I still enjoy the things I used to enjoy'. *Anxiety:* 'Worrying thoughts go through my mind' & "I get a sort of frightened feeling like 'butterflies' in the stomach" contributed a large proportion (21 & 23%) to the group difference. 'I get a sort of frightened feeling as if something awful is about to happen' and 'I can sit at ease and feel relaxed'	*Strengths:* 1. The authors stated that strong anxiety and depression latent variables were observed, and each item demonstrating DIF had relatively little shared variance with the exogenous group variable. 2. The authors highlight other strengths: a) the population-based nature of both groups; b) the "reasonably high" response rate for both groups of women (66% for women with cancer and 56% for the reference group); c) the relatively little pre-selection bias against high anxiety or depression by consenting clinicians for the patient group;

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	**Magnitude & Impact**	
violation of conditional independence of item scores. All items were specified to load on one factor. If improved fit resulted from including a correlation between the residual variance of an item and group membership, the item was considered as having DIF. EQS 5.6 was used. Group differences were examined within each scale (Osborne et al., 2004).	df, p < 0.001). The added correlations were between the group variable and the residuals of "I get a sort of frightened feeling as if something awful is going to happen"; correlation = 0.14; covariance= 0.086; standard error (SE)= 0.017 and "I get a sort of frightened feeling like 'butterflies' in the stomach"; 0.096, -0.069, 0.017. (pg 196) *Depression:* Two depression items ('I still enjoy the things I used to enjoy' and 'I have lost interest in my appearance') had negative DIF. According to the authors, the time reference does not explain why women with breast cancer have a lower score in the direction of depression. Three other depression items, 'I feel cheerful', 'I feel as if I am slowed down', and 'I can enjoy a good book or radio or TV program', had positive DIF but do not refer to time. (pg 202) Items not affected by DIF contain references to the past ('I can laugh at the funny side of things' and 'I look forward with enjoyment to things'). The direction of DIF across items with similar content was not consistent.	contributed very little (7& 3%). The remaining 3 items contributed 13% to 17% to the overall sub-scale difference. *Depression:* 'I still enjoy the things I used to enjoy' and 'I have lost interest in my appearance' contributed almost 2/3 of the scale difference (29& 35%); 'I feel cheerful' contributed little (4%); 'I feel as if I am slowed down' contributed in the opposite direction – 7% (women with breast cancer score higher on this item than the reference) (pg 195-196). *Impact:* "For the anxiety scale, the negative DIF of "I get a sort of frightened feeling like 'butterflies' in the stomach" was 'balanced' by the positive DIF of 'I get a sort of frightened feeling as if something awful is about to happen', resulting in a near zero net effect on the group difference estimate." The correlation between the group variable and the anxiety latent variable was significant (- 0.121). This value might be compared with an estimate of the correlation between the unit-weighted summed score (Rc= 0.83) and the anxiety latent variable of - 0.107. Adjustment for DIF would marginally	d) the fact that participants of the reference group were matched to the patient group by education, country of birth and marital status. 3. Study findings seem to indicate that the lower anxiety and depression levels observed in an earlier study of women with breast cancer as compared to reference group women were confirmed, and were not attributable to DIF in the HADS. The need for replication of findings, perhaps using a different method for examining DIF is highlighted. *Possible Limitations*: 1. The authors caution that the SEM approach used might be lacking in some features provided by the IRT DIF detection method. It is unclear whether or not the method used for assessing DIF was the most appropriate because non-uniform DIF was not examined. 2. The authors suggested that data collection methods might have affected the level of anxiety and/or depression reported by the breast cancer group. 3. No data on non-participants were provided, thus the potential "healthy" self-selection bias could not be discussed. 4. The sample size of the reference group might not have been sufficient for analyses.

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
		increase the excess anxiety in women with breast cancer, compared with population women. The differences between women with breast cancer and reference women on both anxiety and depression were very similar when estimated from the 'DIF adjusted' latent variable as compared with the unadjusted unit-weighted scale scores.	5. DIF was not examined for other variables such as education.
The Patient Health Questionnaire depression scale (PHQ-9). *DIF method:* Exploratory principal components factor analysis was used to derive the factor structure of the PHQ-9 in each of the 4 racial/ethnic groups (non-Hispanic white, African American, Chinese American, and Latino). A generalized Mantel-Haenszel (M-H) statistic was used, which tests for the	*Non-Uniform:* Cannot be tested with the methods used. *Uniform:* 'Abnormalities in sleep' (scores range=0.96 to 1.37) and 'low energy' (scores range=1.24 to 1.41) were the 2 items endorsed most frequently in all groups. 'Depressed mood', 'decreased concentration', and 'thoughts of death or self-harm' were not significantly different between groups. "Chinese Americans endorsed psychomotor abnormalities at a rate more than double the other groups: 0.8 compared with 0.24 to 0.35 (F=104.99, df=3, P<.001). They endorsed abnormalities of appetite at a rate less than half the other groups: 0.4 compared with 0.85 to 0.95 (F=64.10, df=3, P<.001). Chinese Americans also had significantly higher mean scores in abnormalities in sleep	*Magnitude:* Although formal magnitude measures for MIMIC are not available, modification indices can be used as proxy measures of strength and lack of model fit associated with DIF. Most of the DIF found in Chinese American and Latino groups using the M-Hl method no longer showed a significant level of DIF after controlling for covariates of age, sex, and English-language ability in the MIMIC model test. The MIMIC model test indicated that the depressed mood and low energy items did not have significant DIF when controlling for sociodemographic factors. No DIF was found for the African-American group. *Impact:* The authors showed that threshold	*Strengths:* 1. Covariates such as age, sex, and English-language ability were examined in the context of DIF. 2. As per the authors, this study is the first to examine the factor structure of the PHQ-9. 3. Comparisons of the proportions across groups with scores over the threshold is a strength. *Possible Limitations*: 1. The authors stated as a limitation of the Mantel-Haenszel statistic its inability to control for additional covariates beyond depression level. 2. The authors point out that having excluded from analysis subjects who did not endorse at least a screening item (depressed mood, anhedonia, insomnia, or low energy) may have led to reduced variance in

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
equality of odds ratios across several strata. The non-Hispanic white group was used as the referent group in all separate analyses. A Multiple Indicators Multiple Causes (MIMIC) model was used to test for covariates (age, sex, and English language ability) (Huang et al., 2006).	(F=25.71, df=3, P<.001). Latinos had significantly higher mean scores of anhedonia: 0.89 compared with 0.56 to 0.67 (F=23.63, df=3, P<.001). In comparison with Chinese Americans and non-Hispanic whites but not African Americans, Latinos had lower mean scores of abnormalities in sleep (F=25.71, df=3, P<.001), low energy (F=7.85, df=3, P<.001), and guilt (F=11.61, df=3, P<.001). For the Chinese American group, sleep, appetite, and psychomotor changes showed DIF. Anhedonia also had a significant M-H statistic in the Chinese American group, but after controlling for covariates of age, sex, and English-language ability in the MIMIC model test, this item no longer showed significant DIF. For Latinos, anhedonia, sleep changes, appetite changes, and guilt evidenced DIF.	cutoffs derived from the original non-DIF adjusted scores, although significant, were similar, ranging from 15.2% for Chinese Americans to 21.8% for non-Hispanic Whites. However, there were differences between Chinese American males (11.8%) and females (18.1%). "The similar mean scores and factor structure of the PHQ-9 in the different groups suggests that it can be used without adjustment in diverse populations." (pg 550)	response; thus, potentially accounting for the lack of difference in the function and dimensionality of the PHQ-9 between groups. 3. Exploratory factor analysis does not permit examination of factorial invariance. Only weak evidence of dimensional invariance is provided. 4. DIF adjusted estimates of the latent variable means did not appear to be compared to unadjusted means in order to measure impact. 5. It is noted that considerable DIF was observed prior to adjustment for demographic characteristics. Thus DIF in these demographic groups is of concern. Some differences exist in the proportion above threshold.
The mood module of the PRIME-MD (Spitzer, et al., 1994). *DIF Method*: IRTLRDIF was used to evaluate item performance for black and white women. Samejima's graded	*Non-Uniform:* Not found *Uniform:* Not found	*Magnitude:* Item-level DIF was not identified *Impact:* Differential functioning at the scale level was not relevant because no item-level DIF was observed. However in order to examine group mean differences, a multigroup model was estimated using Multilog, in which the item parameters for Black and White women were constrained to be equal. A	*Strengths*: 1. Test for unidimensionality was performed and reported. 2. Findings were replicated with matched samples on marital status, education, housing status, employment status, and insurance status. 3. Anchor item purification was conducted. Items that were not identified as anchors in the first stage were considered candidate DIF items and were evaluated in the second

Measure/ DIF Method	DIF Results		Review
	Non-uniform & Uniform	Magnitude & Impact	
response model was used for item parameter estimation. DIF was identified in a 2-stage approach: selecting anchor items and evaluating candidate items for DIF. For each item a model with estimates constrained to be equal was compared with a model where the item parameters are estimated separately for the groups. This difference is distributed as X^2 with *df* equal to the difference in the number of parameters estimated. The Benjamini-Hochberg stepped procedure was used to adjust multiple comparisons (Hepner et al., 2008).		mean θ score for Black women relative to a mean θ score of zero for White women was estimated using this model. The mean scores for Black women were 0.4 standard deviations lower than that of White women.	stage. 4. Large samples of Black (3,191) and White (315) women were used in analyses. *Possible Limitations*: 1. As noted by the authors, study results are generalizeable to lower income Black and White women only. 2. The sample was from a randomized clinical trial which may also affect generalizeability.

Quality of Life

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
European Organization for Research & Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30) emotional functioning (EF) scale (Aaronson et al., 1993) *DIF Method:* Generalized Partial Credit IRT Model (Muraki, 1992) with 3 thresholds and a slope parameter. “If all items have the same slope, the model is equivalent to the Partial Credit Model (PCM), which belongs to the family of Rasch models. The PCM can be further constrained to the Rating Scale Model...	*Non-Uniform:* Not discussed, although slopes were estimated and examined relative to other items. *Uniform:* The authors did not discuss specific item level DIF, but stated that previous analyses had found no DIF between palliative and non-palliative care groups. They also reported that previous studies had reported identical DIF in the larger QLQ-C30 for language. Thus, most analyses reported here were based on comparisons of language-specific scoring algorithms taking DIF into account or ignoring it. For each language group, ‘Did you feel irritable?’ consistently had the smallest slope and mean threshold, indicating that respondents were less likely to report symptoms based on this item. “Further, ‘Did you worry?’ had the largest mean threshold parameter and the largest or second largest slope.” (Pg 1689) *Item Information Functions (IIFs):* “For the range of IRT scores covering 95% of the palliative care patients (-2.0 to 1.3), ‘Did you feel irritable?’ provided markedly less	*Magnitude:* Not discussed in this article *Impact:* Language-specific scoring algorithms (from separate IRT models for each language group) were compared with a scoring algorithm (from an IRT model based on the total sample). Scoring algorithms that took DIF from language into account did not perform as well as those that ignored DIF. The authors conclude that “when evaluating the impact of cross- language DIF, using a scoring algorithm based on the total sample (i.e. ignoring DIF) gave the best results.” (Pg 1694-1695)	*Strengths:* 1. The IRT analyses presented are thorough, and an excellent discussion of analytic strategies is provided. 2. Detailed analyses of the process used to shorten the scale were provided indicating decisions related to the information provided through the information function. 3. The impact of DIF was examined. *Possible Limitations:* The discussion of the DIF analysis was less well-developed and was based on previous publications. It is recommended that the findings be cross-validated.

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
which can be specified by an item location parameter (e.g. the mean of the thresholds) and category parameters (the difference between the threshold and the location parameters, assumed to be the same for all items with the same response categories)." (Pg 1686) PARSCALE with marginal maximum likelihood was used (Bjorner et al., 2004).	information than the other items. 'Did you worry?' was estimated to be the most informative." "...the IIFs in each of the language groups demonstrated that, for all four groups, 'Did you feel irritable?' provided markedly less information in the range of interest than the other three items." (Pg 1689)		
European Organization for Research & Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30; Aaronson et al., 1993) *DIF Method:* Contingency table of item by studied group	*Non-Uniform:* Cannot be examined with this method. *Uniform:* Three items from three scales indicated significant DIF in both individual cross-validation samples: 1) *Physical Function (PF):* "Do you have to stay in a bed or a chair for most of the day?" evidenced DIF across age and treatment groups 2) *Pain Scale (PS):* "Did pain interfere with your daily activities?" showed DIF across	*Magnitude:* *PF:* "Do you have to stay in a bed or a chair for most of the day?" showed DIF for age: pγ of -0.63 (p= 0.006); older, as contrasted with younger participants were more likely to report having to stay in a bed/chair. The item also showed DIF by treatment group: pγ of 0.74 (p= 0.001); those receiving chemotherapy were more likely to report having to stay in a bed/chair. *PS*: "Did pain interfere with your daily	*Strengths:* 1. This article is a good example of early DIF-detection studies. 2. There is a detailed discussion of the findings and implications. 3. Discussion of magnitude and attempts to examine impact of DIF are strengths of the analyses. *Possible Limitations:* 1. The model does not permit examination of non-uniform DIF. 2. The use of an observed conditioning

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
controlling for score level. Partial gamma ($p\gamma$) was used (Goodman & Kruskal, 1954). Cross-validation was used to control for multiple comparisons. "The test for no item bias was performed by testing for $p\gamma$ equal to zero." The two-sided test probability was estimated using Monte Carlo approximation, and a specialized computer program. The gamma coefficient was also used as a measure of magnitude (Groenvold et al., 1995).	treatment groups 3) *Cognitive Scale (CS):* "...difficulty remembering" also evidenced DIF across treatment groups *Role Function Scale:* DIF could not be examined due to low item variability.	activities?"; $p\gamma$ of 0.61 ($p< 0.001$); patients in chemotherapy, as contrasted with those who were not, reported that pain interfered. *CS:* 'Have you had difficulty remembering things?'; $p\gamma$ =-0.39 ($p<0.001$); more patients receiving chemotherapy reported problems with memory. (pg 811) *Impact:* According to the authors, deleting items with DIF from the Physical Function Scale did not result in an unbiased index. While not a measure of impact per se, the authors examined the "dilution effect" of item bias, defined as a comparison of conclusions drawn from a total sample and from a smaller subscale. They report that different conclusions about pain would be drawn with use of the score vs. item.	score is a limitation. 3. There was a lack of purification. 4. There are relatively small numbers of items for some constructs.
European Organization for Research & Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30; Aaronson et al., 1993)	*Non-Uniform:* As the authors note, slopes were assumed to be equal so that no tests of non-uniform DIF were performed. *Uniform:* Forty percent of the items showed DIF; compared to Caucasians, five were biased against Filipinos, five favored Filipinos, and	*Magnitude:* Not provided, however the authors discuss possible weights to be assigned to items. DIF in the items related to working and constipation is discussed, and recommendations are made to give greater weight to these items for	*Strengths:* 1. The impact of DIF was examined. 2. The authors provide a comprehensive discussion of implications, recommendations, and limitations. For example, they conclude that, subject to cross-validation, the equivalence of quality of life measures may be in question.

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
DIF Method: 1-parameter (Rasch) model (Rasch, 1960) using the marginal maximum likelihood procedure. IRT models were calculated for each ethnic group using the PARSCALE software application. A Bonferroni correction (0.05/30) was used to adjust for multiple comparisons (Pagano & Gotay, 2005).	three favored Japanese. Items that showed DIF for at least one comparison were the first two physical items, the second role item, fatigue, nausea, constipation, the second emotional and the cognitive items, several social items, the last physical and the financial items. The 'worry' item was significantly less difficult for Hawaiians than Caucasians; however, the "work at job" and "constipation" items were less difficult for Caucasians than Hawaiians. The item "social activities" was more difficult for Caucasians and Japanese than Hawaiians and Filipinos. The authors conclude that poor wording and misinterpretation was not the cause of DIF.	Caucasians. Items related to financial difficulties and nausea should be given less weight, and items related to physical function higher weight for Filipinos. *Impact:* Impact was assessed by two methods. Both involved comparison of the original QLQ-C30 score with that of a DIF adjusted score. The first method used item removal and the scored partial correlations, using regression analysis. Twelve of 30 items demonstrated DIF but it was suggested to retain all items in the scale. The authors compared scores with and without DIF items. Both methods resulted in reduction in differences among ethnic groups on the total score (evidence of impact); however, the authors conclude that DIF may not have an effect on the psychometric properties of the QLQ-C30.	*Possible Limitations:* 1. The authors acknowledge the small sample size (61 Filipinos and 51 Hawaiians), and need for replication with a two-parameter model to detect non-uniform DIF. 2. They also point out that the method used cannot detect DIF if most items are biased. (The authors did not perform purification.) 3. A single higher order factor from an analysis reported in another paper is given as support for unidimensionality. DIF was examined using the summation score of 30 items, which measures multiple domains. Although the goal was to examine global QOL, unidimensionality of items should have been tested. 4. Model fit is not discussed.
European Organization for Research and Treatment of Cancer Quality of Life (EORTC QLQ-C30; Aaronson et al., 1993)	*Non-Uniform:* *LR:* "In three cases the interaction term between translation and Emotional Function (EF) scale score was significant ($p < 0.001$), indicating non uniform DIF" for "Did you worry?", and "Did you feel irritable?" when comparing Norwegian with English; for "Did you feel irritable?" when comparing Spanish	*Magnitude:* "A positive γ reflects that, controlling for scale score, subjects completing the translated version reported higher item scores, i.e. better functioning, than subjects completing the English version." For the contingency method, a γ larger	*Strengths:* 1. The authors view the contingency table method as simple, with a straight-forward interpretation of the γ coefficient. 2. They noted that comparisons performed test the comparability of translations in reference to the English (original) version, rather than equivalence across translations.

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
DIF Method: Two different methods were used: a three-way contingency table (CT) method and logistic regression (LR) performed using SAS. LR was used for the comparison of the two methods. DIF was defined as different symptom thresholds across translations (Petersen et al., 2003).	with English. (pg 379) *Uniform:* *CT method:* “Significant DIF ($|\gamma|>0.30$ and $p < 0.001$) was found in 12 cases involving 7 of the 9 translations. (‘Did you feel tense’ for English vs. Swedish, Spanish; ‘Did you worry’ for English vs. Norwegian, Dutch, French; ‘Did you feel irritable’ for English vs. Norwegian, Spanish, and German; ‘Did you feel depressed’ for English vs. Norwegian, Swedish, German, Finnish). Four of these γs were positive and eight were negative.” DIF tests were conducted controlling for age, gender, site, and stage in 11 of the 12 cases where significant DIF was found. Of the resulting 44 stratified DIF tests only two had a partial γs smaller than 0.20, indicating that the DIF could be attributed to the confounder rather than the translation process. (DIF with respect to the Norwegian translation for “Did you feel irritable?” disappeared when controlling for confounding variables.) (pg 377) *LR:* Significant DIF was found in 10 cases involving 6 translations. None of the findings of uniform DIF related to translations were attributed to the confounders: age, gender, site, and stage. Using both methods, significant (uniform)	than .30 was interpreted as “moderate to large” DIF. For the LR method, an odds ratio outside the interval 0.53–1.89 was interpreted as an indication of ‘moderate to large DIF’. Using the CT method, there were significant indications of cross-language DIF in 12 of the 36 comparisons, involving seven of the nine translations. Nine of the 12 significant findings were also significant when using LR to test for uniform DIF. *Impact:* The mean scores were calculated for the Norwegian and English groups using all four items and omitting ‘worry’. While there were significant differences between the groups on the four item scale containing the item with DIF, the difference between groups was not significant when the item was excluded. Thus, the item with DIF could potentially result in a biased result at the scale level. (see pg 381). The authors argue that if a linguistic explanation can be ruled out, the DIF can be attributed to a cross-cultural difference. Two items, the Norwegian translation of “Did you worry”, and the	3. The role of possible confounders in the context of DIF was discussed. 4. Putative causes of DIF (e.g., linguistic DIF due to conceptual non-equivalence and/or cross-cultural differences) were presented. 5. Impact of scale level DIF is discussed. 6. The use of two methods (contingency tables and LR) provided cross validation. Both methods yielded similar results for uniform DIF, “no material DIF found with LR was overlooked using the contingency table method.” *Possible Limitations:* 1. The authors noted that with the contingency table method, the exogenous variable has to be dichotomous or ordinal, and that the partial γ is a test of uniform DIF only. 2. The authors note that lack of purification could have resulted in pseudo-DIF. 3. An observed, rather than the arguably, theoretically preferred latent variable was used as the conditioning variable. 4. No purification was performed. 5. The number of items in the scale (four) for DIF analyses was small. 6. Some of the sample sizes for different countries were small, (e.g., 43, 70).

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	DIF was found in 9 cases (involving 6 translations).	Swedish translation of "Did you feel depressed" could be due to linguistic differences.	
European Organization for Research & Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ-C30) (Aaronson et al., 1993) *DIF Method:* Ordinal logistic regression analysis (using the proportional odds model) was used to derive a common odds ratio for each translation and for each cultural/ geographic group. Absolute log odds ratios >0.64 with a p<.001 were considered significant uniform DIF. English was the reference group in the translation analyses	*Non-Uniform:* While there were 25 instances of statistically significant non-uniform DIF for language, and 9 for cultural/geographic groups, none had an $R^2 \geq 0.035$. *Uniform:* *PF*: Conditional on disability, Swedes were less likely than English speakers to report "trouble doing strenuous activities". Polish respondents reported fewer problems with "taking a long walk" while Dutch, Norwegian and Turkish respondents reported more difficulties "taking a short walk". Dutch and Spanish were less likely to report needing to "stay in bed or a chair during the day" while Taiwan Chinese, Italian, Norwegian, Polish, Swedish and Turkish respondents were more likely to report needing to "stay in bed or a chair during the day." Turkish respondents reported needing more "help with eating, dressing, washing or using the toilet." For cultural/geographic groups: Eastern European (EEu) respondents reported fewer problems with "taking a long walk" while Scandinavian, North Central European (NCE), Islamic and East Asian (EA) respondents reported more difficulties	*Magnitude:* (See non-uniform DIF and use of odds ratio magnitude cutoffs.) The results for the 2 contributing Scandinavian countries were quite different suggesting linguistic issues were more relevant for the "long walk" item. Often the results for the depression items varied by country indicating that translation issues may have impacted the results. DIF findings among different countries using the same translation were generally similar, except Myanmar was significantly different from other English speaking countries. Results for translations were similar when analyses were re-run using IRT to estimate the traits (Scott et al., 2006b). *Impact:* Not discussed.	*Strengths:* 1. As noted by Crane in his commentary (2006), a large multi-national database allowed for robust analyses across 27 countries and 13 languages. 2. Crane notes that qualitative interviews with bilingual people to assess the equivalence of translations added context to the analyses. 3. Crane notes that in addition to conservative p values, the authors took magnitude of DIF into account when determining significance. 4. A thorough discussion of the possible reasons for the cultural group findings is provided. *Possible Limitations:* 1. The authors recognize that the data for some of the languages came from a relatively small number of studies; therefore, some results may reflect the shared characteristics of these populations. 2. Crane notes that sum scores were used as the conditioning variable. Item response theory may have been more appropriate. However, a re-analysis by the authors using IRT (Scott et al., 2006b) yielded similar

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
and the UK was the reference geographic group. Stata 9.0 was used. Interaction terms (the product of each language or cultural group variable and the overall scale score) were used to test non-uniform DIF. Values were considered significant if $p<.001$ and $R^2 \geq 0.035$. Language data were reanalyzed using IRT-scores as the conditioning variable. (Scott et al., 2006a; 2007; comments by Crane, 2006; response by Scott et al., 2006b)	"taking a short walk". Scandinavian, Eastern European, Islamic and EA respondents were more likely to report needing to "stay in bed or a chair during the day." The Islamic group respondents reported needing more "help with eating, dressing, washing or using the toilet." *EF*: Those using the Polish or Singapore Chinese translations were less likely to report "feeling tense". While respondents using the Norwegian, Turkish or the two Chinese translations were less likely to endorse "Did you worry", Germans were more likely to endorse this item. Dutch or Spanish speakers scored lower on "Did you feel irritable?" Norwegian, Swedish, Polish, and the two Chinese translations were more likely to endorse "did you feel depressed". Similar to the linguistic group findings, EEu and EA were less likely to report "feeling tense". The North Central European group was significantly more likely to endorse "Did you worry". Again, EA, EEu, and Scandinavian groups were more likely to endorse "did you feel depressed". *CF:* "Respondents using the Dutch or Swedish translations reported more 'difficulties with concentrating' than with 'remembering things' compared with English speakers."		results. 3. Several scales had only 2 items perhaps making them inappropriate for DIF analysis. 4. Crane suggests that English may not have been the best reference category and others should have been examined. The authors respond that the instrument was initially developed in English, and all translations were based on this version (Scott et al., 2006b). 5. The authors note that the somewhat arbitrary nature of the cultural groupings allowed for certain well-represented countries or translations to dominate the overall regional results. Groupings were based on the number of available respondents. 6. The results of the geographic/cultural analyses may have resulted from translational or other factors. 7. Qualitative interviews may have been more useful if conducted with bilingual measurement experts. 8. DIF impact on scale scores was not examined. 9. There is some redundancy between the linguistic and cultural/geographic groups and findings.

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	SF: The EEu group and the Danish and German speaking respondents were more likely to endorse the "family life" item; the latter 2 groups, and the Spanish speaking group scored lower on the "social activities" item. *Fatigue*: Spanish speakers scored lower on "did you need to rest?". Italian, Spanish and Turkish respondents were more likely to endorse "feeling weak". Norwegian, Turkish and Taiwan Chinese scored lower on "were you tired?". Scandinavians scored significantly lower on "were you tired?". *Nausea*: Italian and Singapore Chinese respondents were less likely to endorse "have you felt nauseated?" EA respondents were more likely to endorse "have you vomited?" *Pain*: Relative to other language groups, English speakers scored significantly lower on the item, "did pain interfere with your daily activities?" SW Europeans were significantly less likely to endorse "have you had pain?"		
EUROQoL (EQ-5D; Kind, 1996) *DIF Method:* Rasch one parameter IRT model (Rasch,	*Non-Uniform:* Not estimable with this method *Uniform:* The authors used correlational methods to examine invariance. Variation was present in the order and location of some of the items	*Magnitude:* Not provided *Impact:* Not provided	*Strengths:* Translations were conducted according to international guidelines. *Possible Limitations:* 1. The authors acknowledge that the wide variety in the sample sizes across countries (from n=33 in Denmark to n=2,958 in

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
1960). WINSTEPS software (Linacre, 2005) was used to estimate item calibrations (Prieto et al., 2003).	for individual countries; however, the pattern of item calibration was congruent. "Items varied in severity from -2.50 (anxiety/ depression) to 1.92 (self-care). The rank of average EQ-5D item calibration values was similar for all of the countries except Denmark. 'Anxiety/depression' (followed by 'usual activities'), was the easiest item (to endorse) in all countries, while 'mobility' and 'self-care' were the most difficult (to endorse)." (pg 27) In contrast, most Danish subjects rated 'mobility' as easy to endorse.		Germany) may be the reason for non-comparability of the mobility item. 2. Visual inspection of a graph of difficulty (severity) parameters was used to identify DIF. 3. No follow-up measurements were performed for the misfitting items. 4. The authors used a Rasch model for calibrations and correlational methods for examination of invariance. Correlations may not be the best method for such analyses (see Rupp & Zumbo, 2004). 5. Non-uniform DIF was not examined. 6. Magnitude & impact were not discussed.
Rheumatoid Arthritis Quality of Life Scale (De Jong, van der Heijde, McKenna & Whalley, 1997) *DIF Method:* Rasch one parameter IRT model (Rasch, 1960). A Bonferroni correction was made to adjust for multiple comparisons, and the p value was set at 0.006 (Kutlay et al., 2003)	*Non-Uniform:* The authors claim to examine non-uniform DIF, however it is not clear how this was performed using the Rasch model. *Uniform:* One item showed DIF for age. Conditional on quality of life, younger, in contrast to older respondents were less likely to report "difficulty in caring for persons they are close to". *Cross-cultural DIF: Americans vs. Turks* 1) "I find it difficult to walk to the shops" (direction not provided) 2) "I sometimes have problems using the toilet" was more difficult for Turkish respondents. (The authors suggest that this is	*Magnitude:* Not provided *Impact:* Separate items for each language version were created for the four items in the original 26 item scale that showed DIF, resulting in a 34 item scale. Fit of the expanded version is good for the pooled data set, with no items showing misfit. Although not a test of impact per se, this procedure could adjust for DIF.	*Strengths:* 1. Recent guidelines for cross-cultural adaptation of instruments were followed. 2. Model assumptions were checked. *Possible Limitations:* 1. The instrument was not administered uniformly; literate subjects self-completed the questionnaire, while illiterate subjects were administered the questionnaire. 2. The sample size was relatively small (n= 71). 3. The analyses are not well described; it is not clear how non-uniform DIF was examined using the Rasch model.

Measure/ DIF Method	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	due to the squatting toilets used.) 3) “Often gets frustrated” 4) “Feels unable to control situation”		
WHOQOL-BREF (The WHOQOL-Taiwan Group. 2000) *DIF Method*; The multinomial Rasch (random coefficients multinomial logit) model was used. Items with locations greater than or equal to 0.5 logits between comparison groups, estimated by using ConQuest, showed DIF. A Taiwan general population sample was used to evaluate gender, age and education (Wang et al., 2006).	*Non-Uniform:* Not estimated with this method. *Uniform:* Item locations were compared. No DIF for gender was found. Three items showed DIF for age: “How well are you able to get around” – “young” vs. “50+”; “How satisfied are you with your sex life” – “below 40” vs. “older”; “Are you usually able to get the things you like to eat?” – “young” vs. “old adults” Four items showed DIF for education groups: “How healthy is your physical environment?”; “Have you had enough money to meet your needs?”; “How available to you is the information you need in your day-to-day life?”; and “To what extent do you have the opportunity for leisure activities?”	*Magnitude:* 0.5 logit criteria was used to select items. *Impact:* Not examined	*Strengths*: 1. Purification was conducted by removing items with significant DIF and those with unacceptable outfit MnSq. 2. Interpretation of DIF items was provided for some (but not all) items. *Possible Limitations:* 1. Items that demonstrated DIF in any comparisons were deleted but the impact of DIF resulting from comparisons of results with inclusion and exclusion was not estimated. 2. More than one DIF method should be applied given that the 0.5 logit may not be statistically meaningfull, in the context of a sample size of 13,083. 3. Non-uniform DIF was not examined. 4. Definitions of comparison groups were not provided, e.g., actual years of education included in “lower education” and “higher education” groups.

General Health

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
FACT version 3 (Cella, 1994), which consisted of physical well-being (PWB), emotional well-being (EWB), social/family well-being (SFWB) and functional well-being (FWB) scales. *DIF Method*; Combination of IRT and ordinal logistic regression. IRT scores were generated by using Graded Response Model as implemented in PARSCALE. The variables examined were: language (English vs. Spanish); race (African-American vs. white); Hispanic ethnicity (yes/no); education (9 and fewer vs. 10+); literacy (6th grade level + vs. lower	*Non-Uniform:* *PWB:* language: “I have a lack of energy”, “I have nausea”, “I am bothered by side effects of treatments”, “I feel sick”, & “I am forced to spend time in bed”; Hispanic ethnicity: “lack of energy”, “nausea”, “sick”, & “forced to spend time in bed”; education: “nausea” & “spend time in bed”. *SFWB:* language: “My family has accepted my illness”; race: “Family communication about my illness is poor” & “I feel close to my partner”; education: “I feel distant from my friends”, “family has accepted my illness”, “Family communication”, “feel close to my partner”, & “I am satisfied with my sex life”; literacy: “Family communication” & “feel close to partner”; age: “I get emotional support from my family”, “Family communication” & “satisfied with sex life”; gender: “I get support from my friends and neighbors”, “feel close to partner”, & “satisfied with sex life”; mode of administration: “Family communication”. *EWB:* language: “I feel sad”, “I am proud”, “nervous”, “I worry about dying”, & “I worry that my condition will get worse”;	*Magnitude:* Three models were included in the hierarchical ordinal logistic regression. For non-uniform DIF, the log likelihood between models were compared by using a chi-square test. For uniform DIF, the relative difference between the parameters associated with OLR estimates is determined by comparing the ratio of the changed coefficients. Ratios of 1%, 5% and 10% criteria were calculated respectively at the scale levels to estimate the impact of DIF. *Impact:* The authors calculated the difference between each subject’s score when accounting for DIF and scores when ignoring DIF. The score differences were then compared to the established “minimal important differences” to determine whether retaining the DIF would significantly impact final measure with any clinical meaningfulness. Variables that were associated with relevant scale-level differential functioning were:	*Strengths:* 1. Intensive purification was used. 2. Sample size was relatively large (n=1615) *Possible Limitations:* 1. As discussed by the author, the sample was one of convenience. 2. There were three criteria used to determine uniform DIF (1 percent, 5 percent, 10 percent); it is not clear which criterion is the correct one. However, the scale level impact of DIF was similar across cutoff criteria. 3. The direction of the DIF findings was not provided.

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	

than 6th grade); and mode of administration (self- vs. interviewer-administration for high literacy patients).
A convenience sample of patients with cancer or HIV residing in three cities was used (Crane, Gibbons, Narasimhalu, et al., 2007).
	race: “worry about dying” & “worry that condition will get worse”; Hispanic ethnicity: “sad”, “proud”, “nervous”, “worry about dying”, & “worry that my condition will get worse”; education: “sad”, “proud” & “worry about dying”; literacy: “sad” & “proud”; age: “sad” & “proud”; gender: “sad” & “proud”.
	FWB: language: “I am able to work”, “I have accepted my illness”, & “I am enjoying the things I usually do for fun”; race: “I am able to enjoy life” & “I am sleeping well”; Hispanic ethnicity: “accepted illness”; education: “accepted illness”; literacy: “accepted illness”; mode of administration: “able to work”, “accepted illness” & “I am content with the quality of my life right now”.
	Uniform:
	PWB: language: “lack of energy”, “I have pain”, & “sick”; race: “lack of energy” & “bothered by side effects of treatments”; Hispanic ethnicity: “lack of energy”, “pain” & “sick”; education: “lack of energy” & “sick”; literacy: “lack of energy”; gender: “lack of energy” & “pain”.
	SFWB: language: “distant from friends” & “Family communication about my illness is poor”; race: “distant from friends”, “emotional support from family”, “Family
		PWB: race;
		EWB: race, ethnicity, language;
		SFWB: all but gender;
		FWB: race, language, education and mode of administration.
		The impact of DIF was observed across all scales for race; physical well-being was least affected in terms of DIF impact by demographic characteristics.	

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	communication” & “satisfied with sex life”; Hispanic ethnicity: “distant from friends”, “Family communication” & “satisfied with sex life”); education: “Family communication”; literacy: “Family communication”; age: “emotional support from family” & “satisfied with sex life”; gender: “distant from friends”, “emotional support from family”, “Family communication”, “close to partner”, & “satisfied with sex life”. *EWB:* language: “proud” & “worry that condition will get worse”; race: “I am losing hope in the fight against my illness” & “worry that condition will get worse”; Hispanic ethnicity: “proud” & “worry that condition will get worse”; education: “sad”, “proud”, “nervous”, & “worry that condition will get worse”; literacy: “sad”, “proud”, “nervous”; age: “proud”, “losing hope in the fight against illness”, & “nervous”; gender: “sad”, “proud” & “nervous”. *FWB:* language: “able to work”, “My work is fulfilling”, “accepted illness”, “sleeping well”, “enjoying the things I usually do for fun”, & “content with quality of life”; race: “able to work” & “accepted illness”; Hispanic ethnicity: “able to work”, “work is fulfilling”, “accepted illness”, “sleeping		

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	well", & "content with quality of life"; education: "content with quality of life"; literacy: "content with quality of life"; gender: "able to work" & "sleeping well".		
FACT-B in breast cancer patients (Cella, 1997) *DIF Method*: An extension of the Rasch model (Rasch, 1960) for rating scale data using WINSTEPS (Linacre, 2005) provided estimates of: a) ability, b) item location, and c) response thresholds between adjacent response categories in the rating scale (there are *m* – 1 thresholds in an *m*-category scale) (pg 238). The steps were (see pg 240): 1. The combined sample was used to obtain threshold estimates; then separate item location calibrations were	*Non-Uniform:* Not detected with this model. *Uniform:* Significant DIF was found in each scale, and different items were flagged across groups. *Trial Outcome Index (TOI):* "enjoy life" and "feel sexually attractive" were significantly above the 95% CI, indicating that Austrians tend to use more negative responses; "bothered by weight change", "energy", and "arms swollen and/or tender" were significantly below the 95% CI, indicating the tendency for more positive response by Austrians across all kinds of TOI. The correlation between item calibrations was .92. (pg 243, 246) *Social/ Family Well-Being* (*SWB):* Conditional on the well-being measure, US respondents report lower "satisfaction [sex life]" but higher "family communication" than Austrians. There was a moderate correlation between item calibrations (*r*= .71). Little change was shown in fit statistics and item calibrations after anchoring the five items without DIF. (pg	*Magnitude:* Not discussed in article. *Impact:* "Adjustment of group mean scores for DIF affected slightly the magnitude but not the direction of the mean differences between Austrian and U.S. patients." "There were moderate effect sizes (ES) for the differences in the TOI and the SWB scores between groups, with and without adjustment for DIF." (see pg 254) EWB scores were similar across groups, before and after adjustments. For TOI, SWB, and EWB there were high correlations between patient scores before and after adjusting for DIF (*r*=.99). DIF adjustment resulted in significantly better measurement precision. TOI: Austrian group decreased from SE=2.80 to 2.69, the U.S. group decreased from SE=2.81 to 2.68. *TOI*: Austrian participants had slightly higher physical function and well-being scores, than the U.S. sample; ESs were	*Strengths:* 1. According to the authors, "The current study demonstrated the usefulness of the Rasch model for evaluating the cross-cultural equivalence of self-report instruments...The identification of potentially biased items does not invalidate the questionnaire but, rather, provides an empirical test of construct interpretation..." (pg 255) 2. This is a carefully conducted analysis of DIF in three scales of the FACIT. A strength of the study is the inclusion of the analysis of the impact of DIF. 3. Purification was performed. *Possible limitations*: 1. Limited generalizability may result because groups were not equal in sociodemographic and clinical characteristics, and data collection methods varied in the U.S. sample. 2. There were relatively small group sizes. 3. The use of the one-parameter Rasch model (without extension) does not permit examination of non-uniform DIF.

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
obtained for the two groups. 2. Mean locations for each item were plotted and Spearman rank correlations were used to examine the association between item calibrations. 3. An absolute positive value > 1.96 indicated significant DIF. Multiple comparisons were not adjusted. Purification was performed, and the DIF analysis repeated. SAS was used to create plots (Hahn et al., 2005).	247) *Emotional Well-Being (EWB):* All six items fell in the range of good fit. "Proud of coping", "worry [dying]", and "sad" were shown to differ significantly between the U.S. and the Austrian samples. The correlation between item calibrations was .88. Little change was observed in fit and calibrations after anchoring the three items without DIF. (pg 247, 249) DIF of the SWB item, "Family communication about my illness is poor" and of the EWB item, "I am proud of how I'm coping with my illness", was shown to be due to translation problems. Both were reworded in all languages. (pg 253)	.32 vs. .29 (before vs. after) *SWB:* Austrian patients scored lower (ES .34 vs. .36). *EWB:* effect size .02 vs. 06. The differences in effect sizes across the two groups were 1/3 of a standard deviation, a "minimally important difference" (pg 252)	
RAND-36 (Hays, Sherbourne & Mazel, 1993) *DIF Method:* Non-parametric IRT model. Mokken scale analysis for polychotomous items (MSP), an extension of the Guttman scale analysis, which requires	*Non-Uniform:* Not provided. *Uniform:* MSP software: SEARCH (to find unidimensional and cumulative scales) and TEST (to establish scalability and reliability of specific items, comparable to CFA analysis). DIF is examined by counting violations of equal item step order across groups. Across the three disease groups, multiple sclerosis,	*Magnitude:* Not presented. *Impact:* Not discussed.	*Strengths:* Unlike other methodologies, the Mokken scale defines DIF at category levels rather than at item levels. Because DIF in subscales is examined, it may be more appropriate to name it 'differential scale functioning' rather than DIF. 1. The analyses used were carefully performed with attention to the type of missing data; i.e. missing completely at random. 2. A strength is the discussion of factors

Measure/ Source	**DIF Results**		**Review**
	Uniform & Non-uniform	**Magnitude & Impact**	
that as the function increases, so does the latent trait value. According to the authors, while the Rasch model is a logistic model with equal slopes, the Mokken model is less restrictive because the only requirement is that the function increases with the value of the latent trait (See pg 639) (Moorer et al., 2001).	rheumatism, and COPD, there were no indications of DIF. The general health perceptions and vitality subscales, were only moderately strong, according to the authors. Correlation coefficients between the subscales in the RAND-36 ranged from 0.20 to 0.64. *Mental health:* Of 200 checks for violations of the equality of item step order, eight violations were found; two were significant. *General health perception:* Of 128 checks for violations, 30 violated the equality of item step order; 18 were statistically significant.		that may contribute to DIF, such as the frame of reference in items asking for ratings of health in comparison to others (see also Boorsboom et al., 2002 for a discussion of absolute vs. relative forms of bias.) *Possible Limitations:* 1. A possible weakness of this approach is that less information about type of DIF is available. 2. Additionally, the notion of a hierarchical step function underlies the method. This may not be the best approach for a thorough analysis of DIF. The authors state that not much DIF is present in the RAND-36; however, other methods of DIF detection might identify more DIF in this measure.
SF-36 (Ware & Gandek, 1994; Ware & Sherbourne, 1992) *DIF Method:* "The partial gamma coefficient (Goodman & Kruskal, 1954; two-sided α levels of .05) was used to test for a uniform tendency to score higher on the item	*Non-Uniform:* Non-uniform DIF was not discussed; however, related findings were presented. An examination of the three-way table ("health in general" by nation by anchor scale) suggested that the Danish version of the item was less discriminative than the original. Danes with good general health tended to choose the categories, very good and excellent less often than Americans, whereas Danes with poor general health tended to choose the categories, fair and	*Magnitude:* According to an adaptation of the guidelines from the Educational Testing Services (ETS), six items could be classified as B (slight to moderate DIF), and five items could be classified as C (moderate to severe DIF). The classification was in close agreement with the results from statistical tests in the secondary analysis except "I am as healthy as anybody I know", which had significant DIF, but was classified as A	*Strengths:* 1. The authors used a method with few assumptions so that DIF is not confounded with lack of model fit. 2. The use of anchor items is a strength. 3. Examination of magnitude and impact was performed. 4. Cross-validation of results is informative. 5. Extensive qualitative review of the translations was performed. *Possible Limitations:*

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
in one country. Partial gammas (Goodman & Kruskal, 1954) should be 0; a gamma twice its standard deviation indicates significant DIF. Stratified chi-square analyses and inspection of the tables were used to examine non-uniform DIF." Cross-validation was used to correct for multiple comparisons. Data were randomized into two sub-samples (Bjorner et al., 1998).	poor less often than Americans. For the item "I am as healthy as anybody I know", the Danish version was more discriminative than the U.S. version. (pg 1195) *Uniform:* In the physical functioning subscale, significantly negative partial gammas were observed for 'walking… a mile' and 'walk several blocks'; compared with Americans of equal physical functioning, fewer Danes indicated limitations on these items. Compared with equally able Americans, more Danes indicated limitations in "lifting or carrying groceries" and in "bathing or dressing". Given equivalent general health, Danes on average rated their "health in general" poorer than Americans. Analyses were compared to previous analysis using Rasch modeling, in which it was found that, given equal physical disabilities, "walking" was a more difficult item for the U.S. respondents, while "bathing and dressing" were more difficult for the Danish sample.	(no or negligible DIF) because the size of DIF was small (the partial gamma was 0.18 for the total population). (pg 1195) *Impact:* Impact was examined by comparing the full scale to DIF-free reduced versions. The authors concluded that little impact was observed at the scale level, but that use of single items could bias cross-national comparisons.	1. A possible disadvantage is that observed rather than latent variable models are used. The use of observed variables may not be optimal. As pointed out by the authors, latent variable models can perform interval adjustment for DIF. 2. Non-uniform DIF is not discussed in this article. 3. According to the authors, results of contingency table tests of DIF are affected by the sample size and multiple comparisons. "Interpretation of the size of the gamma coefficient, and comparisons, between the gamma coefficient and odds ratio rests on strong assumptions regarding conditional distributions of items and exogenous variables." "The size of the gamma coefficient may overestimate the degree of DIF in items with many response categories."
SF-36 (Ware & Gandek, 1994; Ware & Sherbourne, 1992) *DIF Method:*	*Non-Uniform:* Significance testing was used. *PF*: Older people with poor physical function had less difficulty with "vigorous activities", and "bend/kneel/stoop" than	*Magnitude:* "Limited in kind of work" had a large effect size for gender in the general population (females responded more positively).	*Strengths:* 1. Purification was performed. 2. Analyses were conducted on two large, diverse data sets (the Medical Outcomes Study and the National Survey of

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
Proportional-odds logistic regression was used. An effect size measure (LR-P-DIF) for converting odds ratios to a difference in probabilities between the two groups (Dorans and Holland, 1993) was used. The 1 *df* test of $\beta_3=0$ was used to test non-uniform DIF. (Perkins, Strump, Monahan & McHorney, 2006)	younger people, whereas among those with high physical function, the opposite was observed. In both samples, those with less education and Blacks had less difficulty with “vigorous activities” with more pronounced difference at lower to mid-levels of physical functioning. Among those with low physical functioning, Blacks had less difficulty than Whites “walking more than a mile”, while at higher function levels Blacks had more difficulty than Whites. “Bathing or dressing” showed DIF for age and education in the general population. *Vitality*: Older persons had more pronounced lower scores for “felt full of pep”(both data sets), “had a lot of energy” (general population) and “felt tired” (general population) at the lower to mid range of vitality scores. *Uniform:* *Role – Physical*: Females responded more positively to “limited in kind of work” in the sick population, and older people were more limited in the sick and general populations. *PF*: For “bathing or dressing”, older people had less difficulty than younger in the sick population. *General Health*: In both data sets, “older	*Impact:* Scales were rescored excluding items flagged in both data sets. All group comparisons are similar to the original except that general health perceptions became significantly different across race in the sick sample. As all vitality scale items were flagged, this was not rescored. “The rescored physical function and mental health scales were significantly higher (p<0.001) than the original scales for all comparisons. General health was significantly lower than the original for age groups.” Due to the small number of items in each scale, removal of items is not recommended.	Functional Health Status) in order to allow for broad generalization of results. *Possible Limitations:* 1. According to the authors, covariates that may have explained the DIF may have been excluded from the analyses. 2. The authors note that due to small scale size type 1 error could have been inflated. 3. According to the authors, the fact that data for the sick sample was collected 20 years prior to the analysis may have led to some bias. 4. While aggregate level impact was examined at the scale level, individual level changes were not reported. It is possible that there were salient differences for some people.

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	people were less likely to report ‘getting sick easier’ than others, less likely to ‘expect their health to get worse’, and more likely to rate their ‘health in general’ as worse than younger people.” In the sick population “expect health to get worse” and “health is excellent” showed DIF for gender. Blacks in both datasets were less likely to “expect their health to get worse”. In both datasets, those in the low education group and Blacks scored lower on “health in general”. *Vitality*: Older people had less ‘energy’ in the sick dataset. Older people were less likely to have “felt worn out” in both datasets, and “felt tired” in the sick population. *MH*: In both data sets older respondents and those with less education were more likely to have “felt calm and peaceful”. In addition, the less education group was more likely to report having “been a happy person”. *Additional age findings*: When comparing respondents 55-69 years old with those 18-39 years old, “vigorous activities”, “bend/ kneel/stoop”, “getting sick easier than others”, “general health”, “expect health to get worse”, “had a lot of energy”, “felt worn out”, and “felt calm and peaceful”		

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	showed DIF.		
SF-36 (Ware & Gandek, 1994; Ware & Sherbourne, 1992) *DIF Method:* a 2-factor (measurement & structural) MIMIC model (Muthén, 1984). A no-DIF CFA model incorporating the measurement latent constructs (physical function (PF) and mental health (MH)) and exogenous variables (demographic variables and 5 chronic diseases) was constructed. Backward selection was used. Hochberg's method was used to adjust for multiple tests. (Yu, Yu, & Ahn, 2007)	*Non-Uniform:* Not estimated with this model. *Uniform:* Hypertensives tended to endorse "moderate activity" and "bending/kneeling/stooping" more frequently. "Those with respiratory diseases tended to score lower on vigorous activities. …patients with hypertension were more likely to report more frequency of being nervous, whereas diabetic patients tended to endorse less frequency of "being nervous." Age showed negative effect, ethnic variables showed positive DIF and female gender showed both directions of DIF on PF variables.	*Magnitude:* All item loadings were significant in the CFA model. DIF effects for the chronic diseases were minimal compared with the demographic variables. *Impact:* Not discussed.	*Strengths:* 1. Anchor items were selected and tested iteratively. 2. A large data set permitted examination of multiple demographic and health factors; on the other hand, the sample may not be representative of the general population. *Possible Limitations:* 1. A possible disadvantage is the inability to model non-uniform DIF with this approach. 2. Impact could have been examined through comparisons of unadjusted and adjusted mean scores. 3. The authors note that the ceiling effects in the PF items may limit the power and reduce the sensitivity of the model to detect DIF. 4. The authors note that MIMIC does not allow for examination of domains with few items, resulting in exclusion of several domains. 5. Possible clinical reasons for the observed DIF for medical conditions were not provided.
SF-12 (Ware, Kosinski & Keller, 1996) *DIF Method:* MIMIC	*Non-Uniform:* Not estimated with this model. *Uniform:* DIF coefficients (presented as direct effects	*Magnitude:* Coefficients could be interpreted as a measure of magnitude, but magnitude was not discussed by the authors; only	*Strengths:* 1. The authors point out the relative advantage of the MIMIC approach which allows adjustment for DIF in analyses

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
(Muthén, 1984) was used to simultaneously examine the four exogenous (demographic) variables and the two latent variables (PCS-12, MCS-12). According to the authors, MIMIC was chosen for two reasons: 1. “A dichotomous factor analysis model without exogenous covariates is equivalent to the standard two-parameter IRT model”; 2. Unlike IRT, “in the MIMIC model, multiple exogenous variables and multiple latent variables can be analyzed simultaneously.” (pg 77) Regression analysis was conducted to examine the effect of exogenous variables on the latent variables (Fleishman & Lawrence, 2003).	interpretable as ordinal probit regression coefficients) varied in the extent to which they manifested significant DIF effects. All but 2 items had some degree of significant direct effects (DIF) on the demographic variable being tested. “The role limitation items (physical and emotional) had small DIF effects. Other items such as “felt calm and peaceful”, “felt downhearted”, “have a lot of energy”, and “social activities” had multiple significant DIF effects” (pg 81). “...men, those in older groups and members of minority communities rated themselves more highly than would be expected, based on their underlying physical and mental health” (pg 81). Persons 40 years and over, “report lower functioning in climbing stairs and in moderate activity than would be expected, given their underlying physical and mental health” (pg 82).	significance was presented. *Impact:* The adjusted DIF model for physical health reduced the education differences but resulted in change in gender differences. The magnitude of coefficients for the race/ethnicity indicators increased, but remained non-significant compared with the no-DIF model (pg 81). “The significantly higher mental health effects for Blacks in the no-DIF model, compared with Whites, were eliminated in the DIF model; the magnitude of the coefficient for Black respondents was nearly halved”. “The significantly positive effect for those aged 60 to 69 years was diminished and non-significant, and the positive effect for the 70+ years group became significantly negative when adjusting for DIF” (pg 81). The residual correlation between the physical and mental health factors was 0.596 in the no-DIF and 0.567 in the DIF model, a non-significant result. “After incorporating DIF effects, the 70+ age group had significantly less emotional well-being than the youngest group.”	without item removal. 2. This is a well-executed study. The authors incorporated consideration of design effects due to complex sampling in the interpretations of the results, a practice that is infrequently observed in reports of DIF analyses. 3. The use of the MIMIC model permitted DIF impact assessment through examination of the effects of demographic and clinical variables on physical and mental health in no-DIF models. Findings were: a) Adjusting for DIF in the mental health measure lowered both education and gender differences, and reduced the effects of Black race on mental health. “Prior findings that mental health is higher among Blacks than Whites may reflect DIF” (pg 82). b). “Positive direct effects from the oldest age group to calm, downhearted, and energy were especially large and may contribute to overestimation of mental health among older persons in models that do not adjust for DIF” (pg 82). *Possible Limitations:* 1. A possible disadvantage is the inability to model non-uniform DIF with this approach.

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
			2. The authors report that the final measurement model was factorially complex with three items (general health, energy, and social activities) loading on more than one factor. As pointed out by the authors, large sample sizes can lead to x^2 statistical rejection of well-fitting models.
Sickness Impact Profile (Bergner, Bobbitt, Carter & Gilson, 1981) *DIF Method:* Extension of Rasch's IRT model. Item-level DIF was defined as a difference in item severity parameters that were more than two times the standard error of the trait estimate (Lindeboom et al., 2004).	*Non-Uniform:* The authors appear to have estimated the discrimination parameters, and could have tested for non-uniform DIF. As stated by the authors, most items had an item discrimination ≥2. Three items, "I change position frequently", "I am eating different food/on diet", and "My sexual activity is decreased" were poorly discriminating items. *Uniform:* Twenty-three items were identified with significant DIF. After adjusting for sickness level, younger participants were more likely to disagree with two items ("I get around only by using a walker, crutches", and "I do not walk up or down hills") in the Ambulation category. Those diagnosed with an internal disease were more likely to disagree with four communication items. Items in the Communication category were easier to agree with for victims of stroke than for	*Magnitude:* The authors used a measure of magnitude expressed as standard errors. However, magnitude is not discussed. *Impact:* Not performed.	*Strengths:* 1. The extended Rasch model permitted examination of item discriminations. 2. An appendix provides the formulas for conversion of the raw item scores into weighted sum scores. Each dichotomous item score is multiplied by the item's discrimination parameter and summed across items. 3. A strength is the examination of person-invariance (assessed through examination of the agreement between the equivalent short form and total calibrated forms using Bland-Altman plots). 4. The short forms developed were examined in the context of CAT, and the discussion of this potential application of the SIP is instructive. *Possible Limitations:* 1. Non-uniform DIF is not discussed. 2. Magnitude of DIF was not discussed. 3. Impact of DIF was not formally examined.

Measure/ Source	DIF Results		Review
	Uniform & Non-uniform	Magnitude & Impact	
	those with other diseases (Pg 69). Younger men were more likely to disagree with the mobility item ,"I do not get around in the dark or in unlit places without someone's help."		
Stanford Health Assessment Questionnaire (HAQ; Fries, Spitz, Kraines, & Holman, 1980) *DIF Method:* The authors report that they tested both uniform and non-uniform DIF using the Rasch model. A Bonferroni adjustment was made, and DIF was tested at (p=0.006) (Kucukdeveci et al., 2004).	*Non-Uniform:* Values for non-uniform DIF were shown, but it is not clear how they were estimated; and non-uniform DIF was not discussed. *Uniform:* One item, "grip" showed significant uniform DIF for gender. DIF was evident in the Activities subscale for country; given matched ability, Turkish respondents reported slightly more impairment than did U.S. respondents in overall activities. The individual scoring of each of the three tasks contained in this scale was not reported to be the reason for DIF.	*Magnitude:* Not provided *Impact:* Not provided	*Strengths:* The authors point out the strengths of the use of IRT models to examine DIF. A good description of the motivation for examining the cross-cultural performance of the measure was provided; the need to examine conceptual equivalence is discussed. *Possible Limitations:* It was not possible from the description in the article to determine how non-uniform DIF was examined using the Rasch model. There are entries in a table indicating the significance of tests of uniform and non-uniform DIF after Bonferroni correction; however, it is not immediately evident how these values were obtained.