Differential Item Functioning and Health Assessment

Jeanne Teresi, Ed.D., Ph.D.

Columbia University Stroud Center and Faculty of Medicine,
New York State Psychiatric Institute,
Research Division, Hebrew Home for the Aged at Riverdale

6/1/04

The issue of cultural equivalence in measures is important because of the increasing diversity observed in many societies. The magnitude of disparities in health care among "priority" populations, defined here to include members of racial, ethnic and other socio-demographic groups, has been the focus of recent controversy (Bloche, 2004; Smedley, Stith and Nelson 2003; Steinbrock, 2004). General agreement exists, however, regarding the key findings and conclusions of the National Healthcare Disparities Report (Department of Health and Human Services, Agency for Healthcare Research and Quality, 2003); namely, that disparities in health care delivery do exist, and that standardized measures of health care quality are required. Attempts to develop item banks of health and health care quality that are relatively culture-fair fall under this mandate.

Examination of differential item functioning (DIF) has become central to the investigation of cultural equivalence of measures. Some (e.g., Hunter and Schmidt, 2000) have questioned the utility of examining DIF in "professionally developed measures" in education, arguing that tests of cognitive ability and educational achievement have been shown not to be test biased, producing equivalent performance outcomes, given equivalent test performance.

What is the case for DIF analyses with respect to health-related measures? Elimination of bias in measures of health remains an important goal. It can be assumed that such biases will always exist, and that they are unpredictable because too many factors are at work, and too many cultural background variables exist. An important observation is that, unlike most applications in health, items developed professionally to evaluate educational traits are subjected to extensive substantive qualitative review, followed by DIF analyses, before the tests are finalized and released. Additionally, very large item pools are constructed for such use. However, even in educational testing, a disconnect between substantive and statistical analyses can be observed.

One statistical scholar at Educational Testing Service (ETS) recently commented that, after examining many DIF statistics, someone occasionally remarks, "Maybe we should turn the card (with the item statistics) over and see what the item is." (Paul Holland, February 2002).

Few of the measures used to assess health status have undergone the type of qualitative analyses recommended to examine conceptual equivalence. An important distinction between applications in health, as contrasted with other settings such as educational and aptitude testing, is that there are many health-related constructs and multiple measures of each, few of which have received much critical evaluation. While this situation is beginning to be redressed, with increasing use of Computerized Adaptive Testing (CAT), in which relatively smaller subsets of items are used to establish ability, it is necessary to ensure that the item bank be adequate to the task.

Definitions: The precise definition of DIF varies across methods, and according to whether binary or polytomous (usually ordinal) items are being examined. However, DIF can be defined broadly as conditional probabilities or conditional expected item scores that vary across groups. Controlling for level of health status, is the response to an item related to group membership? For example, a randomly-selected person of color (however defined) with moderate health disorder should have the same chance of responding in the impaired direction to a health status item as would a randomly selected individual also with moderate health disorder, but who is not a person of color. Uniform DIF indicates that the DIF is in the same direction across the entire spectrum of disability, while non-uniform DIF means that an item favors one group at certain disability levels, and other groups at other levels. Non-uniform DIF can be viewed as a significant group by disability interaction, and shown graphically as two item characteristic curves that cross. An example of non-uniform DIF, given by Teresi and colleagues

(2000), is a plot showing that for lower levels of cognitive ability, the probability of responding correctly to a cognitive item is highest for the highest education group; however, at higher levels of ability the advantage is for the lowest education group. Item bias implies that a substantive review has been undertaken, and that the cumulative body of evidence suggests that the item performs differently, may have different meaning or may be measuring an unwanted nuisance factor for one group as contrasted with another. Magnitude of DIF usually refers to the degree of item-level DIF, and is 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. Impact (usually examined at the scale level) implies that there are group differences in the health status measure distributions or in total (test) response functions (reflecting the relationship between the expected scale score and the ability estimate). A broader definition is that DIF has an impact on the relationships of health status variables and predicted outcomes such as access to care, functional decline and morbidity. Typically impact is examined in terms of effect sizes; 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. (See the presentations by Morales and Fleishman for examples.) The examination of predicted outcomes is not frequently examined.

Note that there is considerable controversy about two components of the above definitions. First, a question arises as to whether or not groups defined, for example, by language, race, or ethnicity constitute homogenous meaningful entities or are proxies for other variables and, as such, should be "deconstructed" to include factors such as acculturation, educational background or reading level. Additionally, if groups are to be used, there are

numerous interactions that might be considered. The recent (2004) report of the National Research Council on Measuring Racial Discrimination concludes that while race is a complex social construct, the definition of which is evolving, data on race and ethnicity should continue to be collected and included in policy research.

The second part of the controversy relates to the estimate of the conditioning (matching) variable, defined here as health status or disorder. As a prominent scholar in health disparities research (James Jackson, personal communication, March, 2004) reminds us, the issue is the nature of the conditioning variable. What is θ (theta)? Various methods for examination of differential item functioning have associated techniques for derivation of the estimate of disability, disease, capability, and so forth. At the most elementary level, this is a conditioning total or weighted raw score. Other procedures assume the existence of a latent variable, estimated using marginal maximum likelihood or other procedures. Yet other methods assume a "valid" target dimension that is distinct from secondary "nuisance" factors. Finally, still other methods assume that there is an external "gold standard" diagnostic variable, if one exists, or a "silver standard" "anchor" such as a vignette to describe the target ability (King, Murray, Salomon, Tandon, 2004). An example of this might be five ordered vignettes describing better or worse health behaviors related to food consumption. (It is assumed that the health behavior construct is unidimensional, and that the vignettes represent consistently the level of the construct measured -- they should map ordinally on theta.) Individuals from a random subsample of the targeted population rank each vignette, using the same ordinal categories that are used to assess individual health behaviors. Individual responses to health behavior questions can then be reordered relative to the vignette anchors. A related concept is that there might be two DIF dimensions measured: relative and absolute. A within-group comparison will lead to

absolute bias, but not to relative bias. An example, given by Borsboom and colleagues (2002) is a statement about whether a person would do well on a basketball team. This could show absolute bias for gender groups; because men may view themselves relative to other men, and women, relative to other women; men and women of the same height would not have the same probability of an affirmative answer, although within gender groups they would.

It is important to note that regardless of method, benefit is derived from "purification" of the conditioning measure. This implies that DIF analyses are performed in a staged or iterative manner, with removal from the conditioning variable items found to have DIF in prior stages of the analyses. Thus, for some methods, a set of "anchor" items that do not have significant DIF is identified, and these (together with the studied item) form a purified measure that is used in subsequent stages of DIF analyses.

Different methods: Numerous articles have been written comparing the different methods for examination of DIF (see for example, Camilli and Shepard, 1994; Holland and Wainer; 1993 Millsap and Everson, 1993; Potenza and Dorans, 1995; Thissen, Steinberg and Wainer, 1993). A review of DIF methods that have been used in health and mental health applications can be found in Teresi (2001). The aim here is to provide a very brief orientation to the approaches. Differences among DIF methods can be characterized according to whether they (a) are parametric or non-parametric; (b) are based on latent or observed variables; (c) treat the disability dimension as continuous; (d) can model multiple traits; (e) can detect both uniform and non-uniform DIF; (f) can examine polytomous responses; (g) can include covariates in the model, and whether they (h) must use a categorical studied (group variable). Some of these characteristics are summarized in Table 1.

The simplest way to think about DIF is to envision a contingency table that examines the

cross-tabulation of item response by group membership for every level (or combined levels) of the attribute. Then, there are several DIF statistics that can be developed from this basic cross-tabulation; e.g., Mantel-Haenszel (MH) (Mantel and Haenszel, 1959; Holland and Thayer, 1988); standardization (Dorans and Kulick, 1986); SIBTEST (Shealy and Stout, 1993). Conceptually one can think about DIF as predicting item response from group membership, controlling for ability. This formulation leads to several parametric models that examine the main effects of group, ability and their interaction; the latter term measures non-uniform DIF. Other DIF approaches, anchored in item-response theory (IRT), compare the equality of parameters estimated simultaneously (or equated) for studied groups. In comparison with other approaches to DIF detection, theoretically IRT has many advantages (see Hambleton, Swaminathan and Rogers (1991), however, IRT (and other parametric methods) is based on assumptions, and lack of model fit can be mistaken for DIF.

Several methods are based on examination of likelihood ratios associated with nested models. The group differences in log-likelihoods associated with compact and augmented models are examined, in which the augmented models contain additional terms or parameters, and the compact model is the more parsimonious. A likelihood ratio test, distributed as a chi-square, is examined in order to test the difference between models; if the model fit is better with the augmenting term (if the -2 log-likelihood is smaller and the p value larger), and if the chi-square associated with the difference between the log-likelihoods for two (nested) models is significant, this indicates the presence of DIF.

For example, in logistic regression the hypotheses are based on comparisons of different models that can be tested by likelihood ratio statistics. The aim is to achieve parsimony with the fewest terms represented. There are some variations in approach, depending upon whether

uniform and non-uniform DIF are examined simultaneously; however, simulations have shown superior performance when uniform and non-uniform DIF are examined in separate nested models (see Jodoin and Gierl, 2001). A full (augmented) model that includes all terms is tested, against a reduced (compact) model (that assumes no non-uniform DIF) with the removal of the interaction term. Finally, a further reduced model that assumes no group effect is examined. Operationally, the method is to enter ability, then group, then group * ability. The last step in the analysis represents the augmented (full) model that is tested against a nested model, with only group and ability in the equation. The improvement in the chi-square model fit with 1 degree of freedom from the last step in the model provides a test of non-uniform DIF. Additionally, some methods examine the change in the R^2 at each step (Zumbo, 1999) or in the significance and magnitude of the change in the beta coefficients at each step (Crane and colleagues, 2004).

The IRTLRDIF (Thissen, 2001) approach also contrasts a compact model that assumes equality of parameters between groups with an augmented model, constructed by freeing equality constraints (adding parameter estimates) for the item to be tested. Similar to some applications of the logistic regression approach, the equality of the discrimination parameters is tested prior to those of the difficulty, so that non-uniform DIF is examined first. The models include all items (or all anchor items if these have been identified previously), including the studied item.

Other IRT-based methods use the results of IRT to examine DIF; for example the DFIT methods (Raju and colleagues, 1995; Flowers and colleagues, 1999) can be used in conjunction with the Graded Response Model (GRM) (Samejima, 1969) or other IRT models. Several methods, e.g., standardization, SIBTEST and DFIT) share the concept of an expected item or true score. In the polytomous case, for each item, the expected score is the sum either of the

probability (DFIT) or empirical proportion (non-parametric methods) of scoring in a category, for given levels of IRT-estimated (DFIT) or observed (standardization and SIBTEST) disability. In order to examine DIF in terms of area statistics in the DFIT methodology, the boundary response function (BRF) is used, so that the response curves have the same form across all categories. Group differences in probabilities and expected item scores can be calculated and figure in the magnitude of DIF that is contributed by the item to the overall measure. (Expected item scores are calculated by summing across individuals, the (weighted) probabilities of response associated with the BRFs.) Non-Compensatory DIF reflects differences in the conditional probabilities of response to an item for randomly selected individuals from the focal group and from the reference group. A magnitude measure developed by Raju and colleagues (1995), and modified by Flowers and colleagues (1999), is the compensatory DIF (CDIF) statistic (see also the contribution by Morales). Because differential test functioning (DTF) is a function of CDIF, CDIF can be used to examine how much each item's CDIF contributes to DTF.

The multiple-indicator multiple-cause (MIMIC) model (see Muthén, 2002) is linked to the normal ogive version of IRT as originally proposed by Birnbaum (Lord and Novick, 1968). One distinct feature of this model is that it is also linked to a structural equation modeling framework in which both direct and indirect effects can be measured as path coefficients, so that DIF is defined as a direct effect (β) of the group variable examined (e.g., ethnicity) on the item, after controlling for disability, and other demographic variables, such as education and gender that might affect indirectly item response. In this context only uniform DIF can be examined, here defined as a significant β, or group differences in the location of the item difficulty on the estimate of the latent disability. A strength of this model is its ability to examine and adjust the

impact of DIF during analyses (see the contribution by Fleishman).

Interpretation of DIF: Examination of possible causes of DIF associated with translations, for example, include: changes in content, format, difficulty of words or sentences, and differences in cultural relevance (Allalouf, Hambleton & Sireci, 1999). As summarized by Gierl and Khaliq (2001), substantive reviewers have not been successful in predicting what items will have DIF, and that judgments (after the fact) regarding why DIF occurred also have failed. To reduce this disconnect, Roussos and Stout (1996) recommended several steps that could be taken to merge the methods. The first stage consists of 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. Because this process is largely based on judgment, there may be some error at this step. Substantive reviewers use four sources for the review process: previously published DIF analyses; substantive content considerations and judgment regarding current items; review of archival data -- review of contexts present in other similar data; testing bundles of items according to some organizing principle. The stage two statistical analyses are comprised of confirmatory tests of DIF hypotheses. This type of procedure can be extended to health-related measures through use of qualitative methods.

Steps in examining DIF: Presented below is a possible ordering of steps that could be performed in examining DIF. It is recognized that these may vary, depending upon whether an already existing measure is evaluated or new items are being developed. For example, the following steps are first if a new measure is being developed, however, if an existing measure is being examined, these steps might occur after the DIF analyses, or might not be possible to implement at all. Based on a paradigm developed by Krause (2002), the steps in the qualitative analyses of health measures might be:

1. Convene focus groups to examine how members of different groups perceive the meaning of studied health constructs;
2. Conduct in-depth interviews to supplement information gathered in the focus groups;
3. Convene a panel of experts to review closed-ended items;
4. Perform cognitive interviews, presenting individuals with closed-ended items, followed by open-ended probes.

A frequent goal in measurement evaluation is to examine factorial invariance. While the relationship between DIF and factorial invariance is beyond the scope of this paper, the following practice guidelines are provided as illustrative of their joint use. Because of the complexity of examining factorial invariance, it is probably most useful to perform the quantitative analyses iteratively. Performing a multi-group analysis requires very large sample sizes, so judicious prior model-pruning is advantageous. The following steps might be taken:

1. Examine dimensional invariance separately for each group.

2a. Perform an IRT analyses to test the fit of the models, and select the most appropriate for use with IRT-based or other parametric DIF methods; or

2b. Select a non-parametric DIF detection method, but bear in mind that not all non-parametric methods measure well, non-uniform DIF. (For example, a method such as MH with standardization might be used to pre-screen items.)

3. Perform purification, constructing a valid target or anchor conditioning variable.
4. Depending on the method, use both significance testing and item-level effect size or magnitude measures to determine the importance of DIF.
5. Perform a cross-validation, if possible using a random half of the sample, or an

independent sample.

6. Examine the impact of the DIF in terms of (a) group differences in the total score or latent variable means with and without DIF items included; (b) group differences in the total (test) response functions; (c) differences in relationships of demographic variables to health variables, under conditions of DIF and no-DIF; (d) relationships of the studied measure with criterion variables or outcomes, with and without removal of the offending items.

After all of these analyses have been performed, the worst-offending items could be removed, resulting in a partially purified measure that then could be used to examine the steps involved in testing the hierarchy of factorial invariance. (It is noted that item removal may not be a viable option with shorter measures.) Or, if no further analyses are to be conducted, a DIF adjustment technique could be applied. It should be noted that the above steps will not guarantee configural, metric, intercept and residual measurement invariance, because not all DIF methods require such assumptions, and even if IRT (which does require these assumptions) is used, not all offending items will be removed at this step. However, the DIF analyses provide a starting point that hopefully will result in greater ease in application of the hierarchical tests of factorial invariance, if multi-group models are to be examined.

Practical Considerations: There is sometimes a lack of connection between theoretical research and its application. Thus, the techniques that are most widely used in practice frequently are far behind the developmental research. In part this has to do with the ease or facility of the methods and availability of software. To borrow from Wall Street, we have the institutional traders (high-stakes testers) and the individual investors (investigators) who do not have the resources to use the most state-of-the art methods. However, this is not necessarily bad;

it takes years of debate among psychometricians, and the contributions of numerous simulation studies before the pros and cons of different methods are identified fully. In the case of DIF for example, relatively few applications apply the latest multidimensional models, non-parametric methods and tests of dimensionality. The majority (if not all) of the applications in the health arena are of unidimensional constructs.

What tends to happen, as well, is that there is considerable investment and attendant acrimony resulting from proponents of one method vs. another. Thus, it is likely to be heard, "why did he use that outdated method; why did she use that obviously wrong method". One may ask, are the theoretical differences large in practice?

As an example, there is a body of literature recommending against the use of the MH because it doesn't detect well non-uniform DIF. But why is it (together with standardization) still one of the most used methods, and why is it used routinely at large institutions such as ETS? Investigators, such as Dorans and Kulick would argue that non-uniform DIF can be detected by visual inspection of empirical item-scale regressions, and that non-uniform DIF is relatively rare. (It may be true that non-uniform DIF is rare in measures constructed at ETS because of the long process that goes into item writing, and investigation of items prior to DIF testing.) This is probably not the case with most health-related data (see also McHorney, 2003). But MH does not give a terribly wrong answer; because it typically is used after extensive qualitative analyses, and in conjunction with standardization and magnitude measures, it can be considered as a screen that can identify first-level DIF, and be useful in identifying anchor items. In this context such non-parametric methods are easy to program, easy to understand and easy to use. Nonetheless, it is important for the applied field to advance beyond what is the easiest to use and most familiar because simulations and studies of real data can inform us about the best methods.

As concluded by the author of a recent simulation study comparing different models and methods, practitioners should begin to move away from selection of a DIF method based on personal preference, but select methods that are appropriate to the data analyzed (Bolt, 2002).

Possible Recommendations: Reviewed in the accompanying table are some possible advantages and disadvantages associated with the use of different DIF methods. Rather than to summarize these, some tentative recommendations for use of these methods with health data are provided, recognizing that all recommendations have associated caveats. First, ease of use, although not the only consideration does play a role in the decision regarding the method. Clearly the easiest methods to use are the non-parametric methods (MH and standardization). Somewhat more difficult are SIBTEST, the LR IRT method, and logistic regression (if an IRT estimate is to be used as a conditioning variable.) Other methods (e.g., MIMIC) are more labor-intensive, and involve more steps (e.g., DFIT).

Other considerations in choice of method relate to decisions about the purpose of the analyses, type of data examined and the sample size. For example, model misfit is a concern because even slight misfit can be mistaken for DIF. Thus, if there is doubt about the model fit of the data, a non-parametric method might be considered. On the other hand, simulation studies have shown that with smaller sample sizes, parametric methods have been shown to be more powerful (Bolt, 2002), and may lead to more stable results (see Wainer 1993). Additionally, with shorter tests, methods relying on the observed score rather than on a latent conditioning variable may be less accurate (see Millsap and Everson, 1993). If there is evidence from the requisite initial studies of dimensional invariance that a unidimensional model fits, and that non-uniform DIF is not critical, then the MH and standardization methods might be used; if such is not the case, Crossing SIBTEST offers an alternative, but is more labor-intensive because a valid

target measure is required. However, in the development of a new measure, hopefully such will be achieved through the use of the steps outlined for the qualitative analysis conducted in Stage 1. If there is concern that non-uniform DIF is present, IRTLRDIF could be used to arrive at a purified anchor set that could be tested further, possibly with DFIT because this companion method will provide a measure of the contribution of the item-level DIF to the total test score. However, more investigation regarding the optimal cutoff values for DFIT is needed. If multiple abilities are being assessed, and/or if the studied variable is not a group variable, logistic regression provides an appealing method for DIF detection, particularly if the IRT estimates for ability are substituted for observed score. If the intent is to study relationships among variables, and measure impact of DIF, the MIMIC model is attractive (see also the contribution by Fleishman). Again, it might be useful first to study the items with the LR IRT method in order to determine the extent of non-uniform DIF, and then to apply MIMIC or, as suggested above, multi-sample analyses with programs such as MPLUS (Muthén and Muthén, 2004). (As stated previously, the latter approach is labor intensive and requires large sample sizes.) Additionally, some new approaches such as the use of hierarchical linear models and anchoring vignettes may be of future use.

Future Directions: DIF, CAT, and Impact: An area requiring more research is the use of DIF methods in the context of CAT. For example, SIBTEST has been expanded to accommodate the use of IRT-based matching variables for application in CAT (Nandakumar and Roussos, 2002). IRTLRDIF (Thissen, 2001) is another promising advance to streamline the DIF procedures in concert with CAT. Another area of needed research is that of examining the impact of DIF. Does it make a difference? The answer is that it depends. For example, in the area of cognitive assessment, considerable DIF has been identified across different groups,

defined by race, ethnicity, education and language, but in some cases, the DIF cancels at the test level because some items favor the focal (studied) group and others the reference groups. Several studies have shown DIF cancellation; e.g., Morales, Reise and Hays (2000); Orlando and Marshall (2002). However, other research has identified impact of DIF as evidenced in changes in prevalence rates (Teresi and colleagues, 1989) or in relationships with mental and functional health variables (Fleishman and Lawrence 2003; Fleishman, Spector and Altman. 2002) as a result of adjustment for DIF. (See also McHorney, 2003 for other citations.)

An important point is that just because DIF cancels at the aggregate measure level, does not mean that individuals may not be affected. For example, if individual items are clinically meaningful, and used diagnostically by clinicians, an adverse impact of DIF could result for an individual. Additionally, some research has shown that even though DIF cancels for an entire measure, measures formed from subsets of items showed non-cancelling DIF. This could produce a deleterious outcome in the context of CAT where subsets of items are selected from an item bank. The implications are that more research should be performed examining differential person functioning (e.g., Johanson and Alsmadi, 2002), using person characteristic curves where performance for a person on each item set (the average proportion correct across several difficulty cluster levels) is plotted against the item difficulty group.

In summary, DIF assessment of measures remains an important component of health disparities research, and of efforts to achieve cultural equivalence in an increasingly, culturally diverse society.

References

Allalouf, A., Hambleton, R. & Sireci S (1999). Identifying the causes of translation DIF on verbal items. Journal of Educational Measurement, 36, 185-198.

Bloche MG. Health care disparities – Science, politics, and race. The New England Journal of Medicine. 2004;350:1568-1570.

Bolt DM. A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. Applied Psychological Measurement. 2002;15:113-141.

Borsboom D, Mellenbergh GJ, van Heerden J. Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. Applied Psychological Measurement. 2002; 26, 433-450.

.Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, California: Sage Publications; 1994.

Chang H, Mazzeo J, Roussos L. Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. Journal of Educational Measurement. 1996;33:333-353.

Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004;23:241-256.

Department of Health and Human Services. National Healthcare Disparities Report. Rockville, MD: Agency for Healthcare Research and Quality, July 2003; December 2003.

Dorans NJ, Holland PW. DIF detection and description: Mantel Haenszel and standardization. In Holland PW, Wainer H, eds. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:35-66.

Dorans NJ, Kulick E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of

Educational Measurement. 1986;23:355-368.

Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: true differences or differential item functioning? Medical Care. 2003, 41:Supplement ppIII-75-III-86.

Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. Journal of Gerontology: Social Sciences. 2002;57B:S275-S284.

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999;23:309-326.

Gierl MJ, Khaliq SN. Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. Journal of Educational Measurement. 2001, 38: 164-187.

Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park, California: Sage Publications, Inc; 1991.

Holland PW, Thayer DT. Differential item performance and the Mantel-Haenszel procedure. In: Wainer H, Braun JI,eds. Test Validity. Hillsdale, N.J.: Lawrence Erlbaum. 1988.

Holland PW, Wainer H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum. 1993.

Hunter JE, Schmidt FL. Racial and gender bias in ability and achievement tests. Psychology, Public Policy and Law. 2000;6:151-158.

Johanson G, Alsmadi A. Differential person functioning. Educational and Psychological Measurement. 2002;62:435-443.

King G, Murray CJL, Salomon JA, Tandon A. Enhancing the validity and cross-cultural comparability of measurement in survey research. American Political Science Review. 2004;98:191-207.

Krause N. A comprehensive strategy for developing closed-ended survey items for use in studies of older adults. Journal of Gerontology B Psychological Sciences. 2002;57B:S263-S274.

Li H-H, Stout W. A new procedure for detection of crossing DIF. Psychometrika. 61:647-677.

Lord FM. Applications of item response theory to practical testing problems. Hillsdale New Jersey: Lawrence Erlbaum; 1980.

Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading Massachusetts: Addison-Wesley Publishing Co; 1968.

Mantel N, Haenszel WM. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute. 1959;22:719-748.

McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Annals of Internal Medicine. 2003;139:403-409.

Millsap RE, Everson HT. Methodology Review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17:297-334.

Morales LS, Reise SP, Hays RD. Evaluating the equivalence of health care ratings by whites and hispanics. Medical Care. 2000;38:517-527.

Muthén BO. Beyond SEM: General latent variable modeling. Behaviormetrika. 2002;29:81-117.

Muthén LK, Muthén BO. MPLUS Statistical Analysis with latent variables. 2004; Users Guide. Los Angeles, California: Muthen and Muthen.

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

Nandakumar R, Roussos L. Evaluation of CATSIB procedure in pretest setting. Journal of Educational and Behavioral Statistics, in press.

Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. Psychological Assessment. 2002;14:50-59.

Potenza MT, Dorans NJ DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement. 1995;19:23-37.

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

Rogers HJ, Swaminathan H. A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement. 1993;17:105-116.

Roussos L, Stout W. A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement. 1996;20:355-371.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph Supplement 1; 1969.

Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika. 1993;58:159-194.

Smedley BD, Stith AY, Neslon AR, eds. Unequal treatment: confronting racial and ethnic disparities in health care. 2003;Washington D.C.: National Academies Press.

Steinbrook R. Disparities in health care -- From politics to policy. New England Journal of Medicine. 2004;350:1486-1488.

Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement. 1990;26:361-370.

Teresi JA. Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. Journal of Mental Health and Aging. 2001;7:31-40.

Teresi J, Cross P, Golden R. Some applications of latent trait analysis to the measurement of ADL. Journal of Gerontology: Social Sciences. 1989;44: S196-S204.

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to Cognitive Assessment Measures. Statistics in Medicine. 2000;19:1651-1683.

Thissen D. MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.; 1991.

Thissen D. IRTLRDIF v2.0b;; Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page. 2001.

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

Wainer H. Model-based standardized measurement of an item's differential impact. . In: Holland PW, Wainer H, eds. Differential Item Functioning. Lawrence Erlbaum, Inc., Hillsdale NJ; 1993:123-135.

Whitmore ML, Schumacker, RE. A comparison of logistic regression and analysis of variance differential item functioning detection methods. Educational and Psychological

Measurement. 1999;59:910-927.

Zumbo BD. 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. 1999.

Table 1. Summary of Features of Different DIF Models and Approaches

METHOD	FEATURES			
	MODEL	**DEFINITION AND TESTS OF DIF**	**POSSIBLE ADVANTAGES**	**POSSIBLE DISADVANTAGES**
MANTEL-HAENSZEL (Holland and Thayer, 1988)	Non-parametric; observed, continuous total score Unidimensionality is assumed, but could consider multivariate or propensity score matching (Dorans and Holland, 1993)	DIF is indicated if there is a significant interaction of item by group, controlling for disability level. The test (for binary items) is if the conditional odds of endorsing the item is the same for both groups.	Few model assumptions; Performs favorably in simulations (see Potenza and Dorans, 1995); Provides magnitude measures; Is not labor intensive or complex	No covariates, other than the total score, which is the construct the item purports to measure; Requires group variable; Usually requires collapsing disability into score groups; Non-uniform DIF not detected well More difficult to model multiple attributes; Less powerful in some studies than parametric methods such as logistic regression (Li and Stout, 1996; Rogers and Swaminathan, 1993)
STANDARDIZATION (Dorans and Kulick, 1986)	Non-parametric; observed continuous total score (Could consider multivariate or propensity score matching (Dorans and Holland, 1993)	DIF is the difference in expected performance on an item, given matched score level (For binary items, the difference in average weighted proportions endorsing an item are examined at each score level.) Both significance and magnitude is used to determine DIF;	Few model assumptions; Provides empirical item- scale regressions, so that non-uniform DIF is detected directly from these plots; Comparing plots across score levels allows visual inspection of item by group by score level interactions Provides magnitude measures with guidelines; Is not labor intensive or complex	No covariates other than the total score; Requires group variable; Formal tests of uniform DIF not available; inspection of plots is used; More difficult to model multiple attributes; Less effective with very skewed data, and (along with other observed score methods) may not be optimal with less than 20 items (see Millsap and Everson, 1993)
SIBTEST (Shealy and Stout, 1993; Poly-SIBTEST (Chang, Mazzeo and Roussos, 1996), Crossing SIBTEST (Li and Stout, 1996) CATSIB (Nandakumar and Roussos, 2002)	Non-parametric, model based on the standardization method. Although the theoretical formulation is multidimensional IRT theory, estimation is based on total continuous, observed disability score for a valid, unidimensional submeasure of the target attribute (the matching variable), adjusted for group differences in latent ability distributions (impact). CATSIB conditions using IRT-based estimates.	DIF is viewed as due to the presence of a nuisance factor; and is the difference between the reference and focal groups in the weighted average item/bundle score on a studied submeasure. The weighting is by the proportion of individuals obtaining a valid subtest score; a regression correction is used to adjust the expected item score for bias in the subtest. Crossing DIF is the average weighted difference between the two marginal IRFs.	Non-parametric, so model fit is not an issue in DIF detection; Allows modeling of multidimensional abilities; Provides DIF significance tests and magnitude estimates; Can measure impact by adjusting means; Can detect crossing DIF with crossing SIB; Simulations show superior performance of Poly-SIB(in comparison to IRTLR and DFIT under several IRT models) in terms of false positives when groups have different ability distributions and the correct model is not known (Bolt, 2002)	No covariates; Usually requires a group or categorical variable; Uses an observed "valid" score that may not be easy to construct; Poly SIB can detect only uniform DIF; May not be powerful with smaller sample sizes (Bolt, 2002; Shealy and Stout, 1993)

METHOD	FEATURES			
	MODEL	**DEFINITION AND TESTS OF DIF**	**POSSIBLE ADVANTAGES**	**POSSIBLE DISADVANTAGES**
LOGISTIC REGRESSION (Swaminathan and Rogers, 1990); ORDINAL LOGISTIC REGRESSION (Zumbo, 1999; Crane and colleagues, 2004)	Parametric; Item response can be realization of a latent continuously distributed random variable; continuous ability variable typically based on observed raw score, but could use IRT-based estimate	Uniform DIF is defined as significant group effect, conditional on ability, and non-uniform DIF as group by ability interactions. Likelihood ratio test where the test statistic is -2(log likelihood for the null model - the log likelihood for the augmented model), distributed as Chi-square Chi-square test can be used with effect size estimate	Covariates can be included; Studied variable can be continuous; Can model multiple abilities (that are not colinear); Can model non-uniform DIF; Performs well in simulations (better in terms of detection rates than MH and Rasch logit) in the presence of non-uniform DIF and when the focal and reference groups have unequal ability distributions (Whitmore and Schumacker, 1999); Provides magnitude measure; Easy to perform (unless IRT ability estimates are used)	Requires more model assumptions and is sensitive to misfit; Item scoring may impact DIF detection; Low item variability may result in false DIF detection; and Use of total score as conditioning variable is not optimal (see Millsap and Everson, 1993) , but other estimates can be used (see Camilli and Shepard, 1994; Crane and colleagues, 2004)..
MIMIC Structural Equation Modeling Approach to IRT (Muthén, 1984; Muthén and Muthén, 2004)	Parametric; dichotomous or ordinal item responses; latent continuous ability variable; generalized least squares estimation (Originally based on limited information normal ogive IRT)	Estimate differences in loadings using equality constraints; estimates DIF using direct effects from measurement/SEM model	Simultaneous modeling group differences in the item response and underlying ability; Covariates can be included; Can model multidimensional data; Studied variable can be continuous; Can adjust for impact of DIF	Single group MIMIC model does not handle non-uniform DIF; and no guessing parameter; No direct estimates of person ability; Estimation was based on covariance matrices rather than individual response patterns, however latest version of M-PLUS is no longer based on limited informaiton; Multiple group analyses requires categorical group variables and large sample sizes

METHOD	FEATURES			
	MODEL	**DEFINITION AND TESTS OF DIF**	**POSSIBLE ADVANTAGES**	**POSSIBLE DISADVANTAGES**
IRT Log-likelihood ratio test (IRT-LR) (Thissen, 1991)	Parametric, latent continuous ability variable; several models available; most popular for health data are the logistic and graded response models	DIF occurs if item response functions differ between groups (Lord, 1980). Conditional on disability, the probability of endorsing any one of the item response categories differs across groups. Differences in parameters, area tests, model-based likelihoods can be examined. A comparison of the chi-square associated with the log-likelihood for a compact model is tested against that of an augmented model with parameters for the studied item freed.	Well-developed theoretical models; Can examine uniform and non-uniform DIF No equating required because of simultaneous estimation of group parameters; Can model missing data; Can measure magnitude as differences in expected item scores; Can measure impact of DIF on the total score using total (test) response function (TRF) which shows the relationship between expected scale scores and theta. Simulations show superior performance to non-parametric methods in several comparisons in terms of power, particularly with small samples, e.g., 300 (Bolt, 2002).	Model must fit the data (misfit can result in Type I error inflation – false positive DIF detection); Assumptions must be met; Categorical group variable is required; Magnitude measures not as well-integrated in DIF detection process
IRT-based DFIT (Raju, 1995; Flowers and colleagues, 1999)	Parametric, latent continuous variable	Non-compensatory DIF (NCDIF) are average squared group differences in item "true" or expected raw scores. The expected score is the sum of the (weighted) probabilities of category endorsement, conditional on disability. Differential test functioning (DTF) is defined based on the compensatory DIF (CDIF) index. DTF reflects group differences summed across items	Can examine both uniform and non-uniform DIF, and shares the advantages of IRT models upon which it is based; Magnitude measures used for DIF detection; Impact of DIF on the total score is examined; Can be used with multidimensional IRT models; Simulations (in comparison with IRTLR using several IRT models) showed less Type I error inflation, and more power (in comparison with PolySIBTEST (Bolt, 2002)	Requires parameter equating; Many programs needed: IRT parameter estimation, equating, DIF calculation, graphics for plots of ICCs; While significance tests (Chi-squares) are available for NCDIF and DFIT, they result in false positives for large samples. Thus, cutoff values (similar to goodness-of-fit indices) are used; these require further investigation