HHS Public Access

Author manuscript

Curr Epidemiol Rep. Author manuscript; available in PMC 2020 June 15.

Published in final edited form as:

Curr Epidemiol Rep. 2019 June 15; 6: 174–183. doi:10.1007/s40471-019-00194-5.

Differential item functioning and its relevance to epidemiology

Richard N. Jones, Sc.D

Director, Quantitative Science Program, Professor, Department of Psychiatry & Human Behavior, Department of Neurology, Warren Alpert Medical School, Brown University, Butler Hospital, 345 Blackstone Boulevard, Box G-BH, Providence, Rhode Island 02906-4800

Abstract

Purpose of review—In this review I trace the origins, applications, limitations and future prospects for research on measurement item bias, or differential item functioning (DIF) in the context of health research. DIF arises in the context of using multiple item or symptom health instruments to rate the level of a particular condition, and describes the situation where not all persons at the same level at the same level of the underlying condition have the same probability of endorsing one or more symptoms. The presence of DIF can lead to biased assessment of group differences and confound risk factor and outcomes research.

Recent findings—The epidemiologic literature includes a great many applied, review, and methodological articles focusing on DIF. The preponderance of the literature appears in the areas of health-related quality of life, physical functioning, cognition, and mental health outcomes.

Summary—Epidemiologists and other researchers in the health sciences often rely upon multiple item rating scales or questionnaires to assess for the presence of or level of health conditions or states that are otherwise not directly observable. When population subgroups respond differently to a subset of the items, this is referred to as differential item functioning (DIF), and might be a source of bias.

Keywords

Epidemiology; IRT; DIF; Measurement invariance; Psychometrics; Factor analysis

Introduction

Clinicians and researchers often make use of multiple item rating scales or questionnaires to assess for the presence of, or level of, health conditions or states that are otherwise not

Richard_Jones@Brown.edu.

Conflict of Interest

Richard N. Jones declares no potential conflict of interest.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

directly observable. The process of assigning numbers to observations, and combining information across multiple items into a reduced number of summary statistics is measurement. Measurement procedures vary in complexity, from simple summation of numbers assigned to rating categories (i.e., true score theory or classical test theory) or more sophisticated methods based on statistical or psychometric models (e.g., item response theory). A number of challenges may present to the researcher working with assessment data and an implemented measurement procedure. Members of a particular sub-group (a focal group) may have lower scores than a reference group and such a finding is not predicted by substantive theory or knowledge. Or, decisions made on the basis of a measurement disfavor the focal group relative to the reference group, and this difference is not expected theoretically, and/or may be socially or politically undesirable. Or, measured values do not predict a criterion outcome similarly in the two groups. The researcher may begin to question whether the assessment instrument and/or measurement procedures are biased in one group relative to the other.

For example, consider screening for dementia with a mental status test. Mental status tests may contain items that involve reading, writing, and arithmetic operations. It may be reasonable to presume that persons with less exposure to formal education may not have the same opportunity to develop the skills necessary to perform well on such items. As a consequence, the meaning of low scores on the mental status test are confounded by level of educational attainment, and the magnitude of this bias may be dependent upon the proportion of items in the test that rely upon skills obtained during schooling, and the impact of this bias may depend upon the context of test use.

Researchers in educational assessment and psychology have developed a set of procedures for delving into these matters. In psychology, the procedures are encompassed within a field of inquiry called measurement or factorial invariance. Bontempo and Hofer (2007) provide an excellent review of the topic, and Bauer (2017) provides an excellent overview of the connections of measurement noninvariance and differential item functioning [1,2]. The range of measurement invariance questions that are considered include whether or not we are measuring the same domains with equal fidelity across groups. In educational research settings, the concern is typically focused on a single domain of measurement, and the questions are therefore limited to the fidelity of measurement of a single domain across groups. Fidelity is here used as a general term with no technical meaning, and is intended to imply a range of possible causes of group differences in the way people are assessed.

Because the research paradigms for investigating measurement differences in assessment instruments have a developmental history in psychology and in educational research, the language used and technical terms borrow from both traditions. For the uninitiated it can be a challenge to navigate the literature and recognize the terms, identify appropriate and useful procedures, and make sense of the research results owing to this variety of approaches. This is compounded by two popular sets of descriptors for levels of invariance in the psychological tradition, one due to Meredith (1993) [3] and another due to Vandenberg & Lance (2000) [4]. For example, in preparing this review I examined the literature for recent publications relevant to differential item functioning in epidemiologic, public health, and general health related journals and identified a number of synonyms used to describe what I

at least consider to be broadly equivalent concepts (Box A). Kim and Smith (2017) blend the psychological and educational traditions in an admirable way in their recent work examining measurement noninvariance in a depression measure among cancer patients [5].

Differential Item Functioning

Differential item functioning is statistical evidence that a particular item performs differently in a focal group relative to a reference group. It is a term developed specifically to separate vernacular notions of item bias from statistical evidence [6]. DIF describes a statistical finding, and the more substantively interesting finding of *item bias* is reserved for instances of DIF where content experts adjudged the source of the DIF to be irrelevant to the construct being measured. It is recognized that the mere presence of a group mean difference on a particular item is insufficient evidence for DIF, because of the possibility that group mean differences in the level of the underlying construct being assessed exist. DIF was therefore operationalized as statistical evidence of a main group level effect conditional on the underlying level of the construct being measured [7].

There are a wide range of procedures described for obtaining such statistical evidence [8]. Millsap & Everson (1993) [9] reviewed many of these methods in use at the time of publication, and provided very strong evidence and a very strong recommendation for using what they termed “unobserved conditioning” methods versus “observed conditioning” methods. Briefly, what Millsap & Everson recommend is that a latent variable method be used to condition the level of the underlying construct being measured when attempting to detect group level effects on a particular item, rather than an observed variable. This implies that latent variable statistical models, rather than statistical models using observed variables (such as sum scores on a test) are preferred for detecting measurement bias. The reason is that attempts to detect DIF on a particular item when conditioning on a total test score suffer from Simpson’s Paradox, because potentially biased items are included in the observed total test score. On the other hand, unobserved conditioning methods, or latent variable approaches including those based in Item Response Theory, are not so limited. One of the pioneers in the field of Item Response Theory, Darrell Bock, commented that if the number of items in a test is very large, the use of observed variables (e.g., sum scores) as a conditioning variable may be appropriate [10]. However, situations in which a large number of items are available is more often encountered in educational assessment and achievement testing, and not so common in health research settings.

In the epidemiologic context, DIF can be viewed as a situation of confounding. DIF occurs when a grouping variable is a confounder of the relationship of the underlying construct and the item response. In Figure 1, I present a directed acyclic graph (DAG) illustrating an instance of DIF. An observed predictor variable x denoting a grouping variable (e.g., sex) causes differences in the underlying construct of interest (f, e.g., depression) and also, independent and in addition to the effect of the predictor in the underlying construct, is related to one particular item on the assessment instrument (y, e.g., tearfulness or crying). Finding that there was a direct effect of sex in tearfulness, independent of the effect of sex in underlying depression and of the effect of underlying depression in tearfulness, is sufficient evidence for concluding DIF exists with respect to tearfulness and sex. Describing the effect

as one indicating item bias would require articulating that cultural standards discourage men, on average or as a group, from exhibiting or admitting to tearfulness or crying behavior.

It is worth pointing out that the heuristic of conceptualizing DIF as confounding should not set limits on the potential meaning and scope of DIF. DIF can also be thought of as or appear as effect modification if the relationship between the item and the latent trait differs according to levels of a third variable (referred to as non-uniform or crossing DIF). The third variable that disrupts the measurement of U via y need not be categorical, and indeed DIF can be more generally thought of as multidimensionality [11]. That is, variability in the item response is attributable to more than a single underlying and potentially unobserved variable.

Item Response Theory

Due to Millsap & Everson's (1993) [9] admonition regarding observed variable conditioning methods for detecting DIF, I will focus on DIF detection methods based on conditioning the group level difference for a unobserved or latent variable. The most important of these for DIF detection work is that of item response theory (IRT) [12]. IRT encompasses a wide variety of statistical models, which vary in their applicability for response variables of different scales (binary, ordinal, continuous) and character of the underlying latent variable(s). I will describe the general aspects of theory for binary variables with a single and presumed continuous normal underlying latent variable, but the reader should recognize that there are many extensions of this basic model [13,14].

The fundamental conceptual unit of IRT is the item characteristic curve (ICC, also known as the item response function or trace line). The ICC is a cumulative probability plot recording the increasing probability of endorsing a symptom (or responding correctly to an individual item in a testing situation) as the level of the underlying trait (θ) presumed to be measured by the symptom increases (e.g., depression, cognitive ability). The functional form of the ICC is presumed to be cumulative normal. Figure 2 displays a hypothetical ICC. The scale of the underlying trait (θ) reflects the customary assumption that the latent dimension is distributed normally with mean zero and unit variance. The ICC embodies several assumptions of IRT; in fact, the normal ogive shape of the ICC is one of the assumptions of IRT [15].

IRT assumes that a latent, unobservable ability or trait describes the probability of responding correctly to an item, and that persons with more ability have a greater likelihood of responding correctly to the item (or endorsing the symptom as the case may be). Many, but not all, IRT models assume this latent trait is continuous normal [16]. Other assumptions of IRT include the notion that the probability of responding correctly to an item, conditional on level of the underlying trait, is independent from responses to other items on the test (conditional or local independence) [15].

Another key notion in IRT is the conceptual and statistical separation of characteristics of persons and characteristic symptoms (or test items). This can be represented symbolically with the item response function

$$P(y_{ij} = 1|\theta_i) = P_i(\theta) = F\{a_j(\theta - b_j)\} \quad (1)$$

which states the probability of the i^{th} individual endorsing the j^{th} symptom (y_{ij}=1 vs. y_{ij}=0) is a function of the discrepancy between the person's level on the underlying latent trait (θ_i) and the level of difficulty or severity measured by the symptom (b_j). This function is modified by the strength of association between the symptom and the underlying trait, the so-called discrimination parameter (a_j). Discrimination refers to the ability of the item to identify persons of high versus low levels on the trait. Highly discriminating items describe steeper item response functions and are better able to separate persons. Parameters a_j and b_j are characteristics of the item, and θ_i is a characteristic of an individual. If an item is ordinal, the probability that a person responds in category c or higher can be expressed as $P(y_{ij} \geq c|\theta_i) = F\{a_j(\theta - b_{cj})\}$ where b_{cj} is a threshold for item j specific to the boundary between category c and $c-1$.

The parameterization of the item response theory function (equation 1) represents a linear model (within a non-linear transformation given by F) of the form $[a(\theta - b)]$. This is the equation of a line in point-slope form. In epidemiology we more often encounter linear equations in intercept slope form, e.g., $P(y_{ij}=1|\theta_i) = F\{B_0 + B_1\theta\}$, where B_0 would represent a constant or intercept term, and B_1 the slope with respect to underlying ability. A little algebra demystifies the point-slope IRT form when we realize that the slopes in the two formulations are equivalent, with $a = B_1$ and $b = -B_1/B_0$. It is also worth noting that the IRT formulation sometimes includes a constant transformation $[Da(\theta - b)]$ where D is scaling constant, the use of which was motivated by a desire to make parameter estimated under a probit and logit transformation comparable [17].

The functional link between the item and person parameters and the response probability [F(•)] are conventionally the logit link or the normal probability link. Historically, the logistic form of the item response function was adopted due to the intense computational demands of the normal probability model [17]. Modern estimation algorithms and advances in computer processing speed have made estimation under the normal probability model a viable option for item response curve analysis. Although in the not so distant past estimation of IRT model parameters required specialized software, procedures are now available in major general purpose statistical software packages including Stata [18], SAS [19] and several packages are available in the R programming environment [20]. Item response theory is also a general form of a more general structural equations model (SEM), specifically a confirmatory factor analysis, as long as the non-normal nature of the response variables are properly modeled [21–23]. Therefore, software for conducting SEM can also be used to estimate the parameters of an IRT model (e.g., Mplus [24], Stata/SEM, R/lavaan [25], among others). Unfortunately, as will be described in greater detail, procedures for identifying DIF are under-developed in general purpose statistical software packages, and applied researchers must use specialty software or user-written programs and algorithms.

DIF in the context of IRT

Simply, DIF in the context of IRT implies that an item response function is different with respect to a third variable, usually a variable indicating group membership. An example is shown in Figure 3. In Figure 3, I have drawn two item characteristic curves for a single binary item, one for a reference group and another for a focal group. The lines are non-overlapping, and the distance between the two lines is proportional across the range of the underlying latent trait. If the lines were plotted on a logit-transformed y-axis, they would be parallel. This is an example of what is called uniform DIF. This is kind of DIF that is illustrated in Figure 1 as well. The grouping variable influences the item response independent of (and uniformly across levels of) the underlying latent trait. This kind of DIF is also referred to as b-DIF, because it is the *b* parameter (equation 1) that is different between the reference and focal group.

In the case of uniform DIF and a binary test item (Figure 3), we see that a randomly selected person from the focal group would need to have a higher level of the underlying latent trait (θ) in order to have a 50% probability of endorsing the symptom than would a randomly selected person from the reference group. We can also appreciate that if we were to use the number of symptoms endorsed as an indicator of a person's level on the underlying trait (e.g., a classical test theory sum score) we would underestimate the level of the latent trait for members of the focal group.

In Figure 4, I illustrate what is called non-uniform or crossing DIF: DIF that is non-proportional across levels of the underlying latent trait. If the lines in Figure 4 were plotted on a logit-transformed y-axis, the lines would be non-parallel. The ICCs in Figure 4 display both b-DIF (the difficulty parameters are different across group) and a-DIF (the discrimination, or *a*, parameters are different across group). Non-uniform DIF is sometimes called crossing DIF because the ICCs, being non-parallel, will cross at some point along the ability axis. Because the item discrimination parameter can be seen as a transformation of the correlation between the latent trait and the propensity to endorse the item [22], we can say that there is a group difference in the factor loading of the item illustrated in Figure 4, being lower in the focal group than in the reference group.

In the case of non-uniform DIF and a binary test item (Figure 4), we see that a randomly selected person from the focal group would need to have a higher level of the underlying latent trait (θ) in order to have a 50% probability of endorsing the symptom than would a randomly selected person from the reference group. We can also appreciate that while *on average* if we were to use the number of symptoms endorsed as an indicator of a person's level on the underlying trait (e.g., a classical test theory sum score) we would underestimate the level of the latent trait for members of the focal group, the degree of underestimation varies as a function of the underlying level of the latent trait.

There is no consensus in the field as to how to quantify the magnitude of DIF. Teresi (2006) [26] reviews these and identifies a broad class of *magnitude* measures that focus on the size of the parameter estimating a group effect or the areas between two group item characteristic curves, and *impact* measures that describe differences on total trait estimate or scale score

that are due to DIF. Crane and colleagues [27] remind us that it can be useful to examine the impact of DIF with graphical methods such as boxplots illustrating the spread of individual differences in latent trait estimates with and without accommodation of detected DIF. This strategy can be especially useful when the latent trait estimate or test score is used in a high stakes situation such as case identification. Hays and colleagues offer a nice example of methods for evaluating DIF impact [28]. Finally, Verdam and colleagues extend this discussion to general cases involving items and third variables confounding measurement to ordinal and continuous cases, and to the evaluation of measurement bias in multidimensional latent constructions [29].

Methods for the detection of DIF

Unfortunately, there exists a great flexibility in the designs, definitions, and analytical modes set against the detection of DIF, and this situation requires one to read the DIF literature with a considerable degree of skepticism [30]. Manuscripts describing new methodological advancements in DIF detection are quite common, and perhaps this is something to be expected in an evolving field of inquiry.

Recently, Teresi and Jones (2013) [8] compiled an exhaustive review of measurement bias and differential item functioning, including a review of software tools used to identify DIF detailed in an appendix. Less attention was given to the algorithm used to identify DIF using those tools. By algorithm, I am referring to the data management decisions, constraints set on the analysis and inference steps, and the order of operations used to carry out a DIF detection procedure. The algorithm used in applied research is rarely specified in sufficient detail to allow replication. In the absence of simulation studies that compare complete algorithms, the applied researcher is left with little guidance on optimal methods.

As an illustration of the potential for differential results with various software approaches and detection algorithms, Yang and colleagues (2011) [31] reported on the parallel detection of DIF in a single data set by different teams using different IRT-based DIF detection procedures. Four different algorithms encompassing different software combinations and algorithm implementations were studied. Evaluated procedures included Item Response Theory Likelihood Ratio Differential Item Functioning (IRTLRDIF/MULTILOG [32,33]), Differential Functioning of Items and Tests/MULTILOG (DFIT/MULTILOG [32,34]), DIFwithPAR/PARSCALE [35,36], and the Mplus/Multiple Indicator, Multiple Causes model [37]. Yang and colleagues found that agreement as to which items displayed DIF was far from perfect, with kappa coefficients ranging from 0.43 to 0.84 across the different algorithm pairs.

It may be surprising that, given a common general understanding of DIF (unequal item characteristic curves) and a common general analytic framework (item response theory) that discrepancies as great as observed in Yang et al (2011) [31] would be observed. But the implementation of DIF detection algorithms is not a standardized set of procedures. In her editorial summarizing a set of articles also approaching DIF in a common data set that appeared in a special issue of *Medical Care* in 2006, Teresi identified many sources of potential discrepancy [26]. Identified elements included dimensionality (the number and

structure of the underlying trait being evaluated), purification (defined below), types and kind of DIF identified (uniform, non-uniform), availability of effect size statistics (magnitude, impact). I expand upon the list of important analytic considerations that face the analyst when embarking upon a DIF analysis in Table 1.

A good number of simulation studies have been published that compare the accuracy of different methods [38–45]. While this is clearly a high priority area for future research, again the applied researcher must be very cautious in reading the methodological literature for clues as to which of the available methods may be advantageous for the context of the planned analysis. The need for caution is underscored by the variety of considerations listed in Table 1, as well as for all those considerations omitted. What is critical to evaluate is the accuracy of a proposed *algorithm* for detecting DIF. This algorithm includes the IRT estimation approach, the method used to detect DIF, and all of the analytic decisions that go along with detecting DIF: the choice or method of anchor item selection, the use of multiple comparisons correction, the statistical criteria used to flag items with DIF, the use of a purification procedure, the procedure (e.g., forwards, backwards) for working through an item set, etc. Readers must scrutinize published simulation studies carefully before using such studies to choose methodologies.

What to do about DIF

The main implication of DIF is that, when detected, sum scores on the instrument are not comparable across group. Dropping items that demonstrate DIF is one extreme way to handle DIF. This is often unpalatable in epidemiologic and other health research settings, as our instruments are relatively short and the offending item may assess a relatively important aspect of the construct of interest. If dropping the item with DIF is not an option, separate calibration with IRT methods is an acceptable way to address the DIF. Now that IRT estimation procedures are available in general purpose statistical software packages, this is probably the preferred option.

Another option is to conduct the substantive analysis within a general latent variable modeling framework [46]. Analysts may include direct effects, group-specific thresholds, or group-by-latent variable interactions [45] as indicated by a DIF analysis, and proceed to address the substantive research question within a general latent variable modeling framework. The advantage of this approach is the population parameters of interest can be estimated while modeling unobservable constructs as latent variables, which should provide more accurate estimate of standard errors and higher quality inferences. Methods that involve two steps (estimating a trait value, and treating it as fixed in a second substantively motivated regression analysis) might return overly optimistic standard errors.

Limits of addressing DIF

It is important to realize that even after removing items with DIF or separately calibrating in two groups to address DIF, this is no guarantee that all undesirable measurement properties of the measurement instrument will be resolved. An example previously mentioned is screening for dementia with a mental status test. Education level DIF has been described for

mental status test items that involve reading, writing, and arithmetic operations [47]. Lower education is also associated with poorer predictive performance of mental status tests for clinical ratings of dementia (higher sensitivity, lower specificity relative to higher education groups [48]). Removing or separately calibrating items that have DIF may not improve the predictive validity of the mental status screening test for clinical dementia, and might make things worse. It depends on whether dementia is truly more prevalent among persons with low education, and depends upon whether the DIF correction has a large impact around the area of the underlying trait that is relevant to the decision-making point on screening, and the adequacy with which the relationship between education and dementia is properly specified in the predictive model.

Future directions and recommendations for applied researchers

There are many exciting new developments in the theory of DIF and in DIF detection procedures. Chief among these may be the use of hybrid models that use latent class or mixture models along with more traditional multiple group or covariate based DIF detection [49]. A motivation here could be that all members of a sub-sample (e.g., males) may not be equally subject to the same cultural influences that impact item response probabilities (e.g., crying, tearfulness). Research along these lines may open up opportunities for DIF detection and amelioration to have increased relevance to individual level inference. Other exciting innovations include the recent generality of moderated nonlinear factor analysis (MNLFA) for DIF detection [2].

For the applied researcher, I will conclude with the following recommendation set for engaging in data analysis activities with the goal of evaluating an assessment method for DIF. The steps are little more than a request for responsible data analysis, with clear specification of methods and procedures to assure readers of the accuracy and statistical power of the algorithm used:

1. **Reproducible data analysis**. As is consistent with movements in epidemiology and other fields [50], the data analysis that performs the DIF detection should be based within a reproducible workflow. Data analysis is at the stage where a reproducible work flow is a core feature of the responsible conduct of research.

2. **Algorithm specification** and ideally, automation. Implementing a DIF detection algorithm can include iterative data analysis steps and multiple decision points. The decision points begin before IRT and DIF detection analysis begins, in the form of decisions for dealing with sparse data (e.g., empty response categories in one or more groups). During DIF detection, reams of data analysis output may be generated. A robust algorithm will at least document and at best automate the processing of handling variables, gathering interim results, and decision making regarding the presence and absence of DIF. This is an essential component of a reproducible workflow. In our own group we have developed tools to facilitate this process in the form of Stata modules[1] governing Mplus MIMIC and multiple group confirmatory factor analysis approaches to DIF detection.

3. **Contextualized algorithm validation**. Before engaging in the primary substantive analysis, data analysts should evaluate their planned algorithm for accuracy (or type-I error level) and statistical power in a Monte Carlo simulation framework. The simulation framework should match the planned analysis in terms of the sample size, number of items, expected number of items with DIF, expected magnitude of DIF. This is exactly the approach one would use to determine adequate sample size to conduct an analysis and test a hypothesis in a study design context. However, in the context of DIF detection the stakes are higher. DIF analyses are usually exploratory, as the causes and direction of differential item functioning are challenging for experts to identify *a priori.* The precise algorithm a data analyst may wish to use with the kind of outcome measure planned in the target analysis. To avoid proposing to modify assessment tools on the basis of results obtained using untested procedures in inadequate samples, DIF analysts need to provide themselves and their audience assurance that the methods employed are robust and appropriate for the target sample. In this digital age long and detailed appendices (or even better, public software repositories) provide a platform to easily share code and advance the field through dissemination of good procedures. Reviewers and editors should insist on this evidence.

Conclusion

The characterization, detection, and amelioration of bias is a fundamental activity for epidemiology [51]. Measurement non-invariance and differential item functioning is an important source of systematic measurement error that can lead to improper inferences and bad judgements. The rapid pace of developments in computer hardware and software on the one hand put the tools necessary to identify measurement non-invariance into the hands of applied researchers, but on the other hand offer a bewildering array of options with no widely accepted and stereotyped modes of analysis [30] which can lead to incomparable results. By embracing behaviors consistent with the responsible conduct of quantitative research, applied epidemiologists have the opportunity to advance the science of measurement in health research and beyond.

References

1. Bontempo D, Hofer S. Assessing factorial invariance in cross-sectional and longitudinal studies In: Ong A, van Dulmen M, editors. Handbook of methods in positive psychology. Oxford University Press; 2007 p. 153–75.
2. Bauer DJ. A more general model for testing measurement invariance and differential item functioning. Psychological methods. 2017;22(3):507. [PubMed: 27266798]
3. Meredith W Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993;58(4):525–43.
4. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods. 2000;3(1):4.

[1]Interested readers can type “net from http://s3.amazonaws.com/mplusmimicbucket” and install our Stata module mplusmimic, which automates Mplus/MIMIC and multiple group confirmatory factor analysis DIF detection algorithm.

•5. Kim J, Smith T. Exploring measurement invariance by gender in the profile of mood states depression subscale among cancer survivors. Qual Life Res. 2017;26(1):171–75. [PubMed: 27837383] Kim and Smith provide a nice example of blending the measurement invariance and differential item functioning modes of analysis.

6. Cole NS. History and development of DIF In: Holland P, Wainer H, editors. Differential item functioning. New York: Routledge; 1993 p. 25–29.

7. Camilli G, Shepard LA. Methods for Identifying Biased Test Items. Sage Publishers: Newbury Park, California; 1994.

8. Teresi JA, Jones RN. Bias in psychological assessment and other measures In: Geisinger K, Bracken B, Carlson J, Hansen J-I, Kuncel N, Reise S, et al., editors. APA handbook of testing and assessment in psychology, Vol 1: Test theory and testing and assessment in industrial and organizational psychology APA handbooks in psychology. Washington, DC: American Psychological Association; 2013 p. 139–64.

9. Millsap R, Everson H. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17(4):297–334.

10. Bock DR. Different DIFs: Comment on the Papers Read by Neil Dorans and David Thissen In: Holland P, Wainer H, editors. Differential item functioning. New York: Routledge; 1993 p. 115–22.

11. Oort F Using restricted factor analysis to detect item bias. Methodika. 1992;6:150–66.

12. Lord F, Novick M. Statistical theories of mental test scores. Addison-Wesley: Reading, MA; 1968.

13. Embretson SE, Reise SP. Item Response Theory for psychologists. Lawrence Erlbaum Associates: Mahwah, New Jersey; 2000.

14. Reckase MD. Multidimensional Item Response Theory. Springer: New York; 2009.

15. Hambleton RK, Swaminathan H, Rogers H. Fundamentals of item response theory. SAGE Publications: Newbury Park; 1991.

16. Woods CM. Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. Psychological methods. 2006;11(3):253. [PubMed: 16953704]

17. Camilli G Teacher's corner: Origin of the scaling constant D=1.7 in item response theory. Journal of Educational and Behavioral Statistics. 1994;19(3):293.

18. Raykov T, Marcoulides GA. A course in item response theory and modeling with Stata. Stata Press: College Station, TX; 2018.

19. Matlock Cole K, Paek I. PROC IRT: A SAS procedure for item response theory. Applied psychological measurement. 2017;41(4):311–20.

20. Rusch T, Mair P, Hatzinger R. in Discussion Paper Series of the Center for Empirical Research Methods (ed Regina Dittrich ML, Katharina Miko, Thomas Rusch, Michael Schiffinger) (Center for Empirical Research Methods, WU Vienna, Austria, Vienna, 2013). http://epub.wu.ac.at/id/eprint/4010

21. Takane Y, De Leeuw J. On the relationship between item response theory and factor analysis of descretized variables. Psychometrika. 1987;52(3):393–408.

22. Lord F, Novick M. Latent traits and item characteristic functions (Chapter 16) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley; 1968 p. 358–93.

23. Mislevy RJ. Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics. 1986;11(1):3–31.

24. Macintosh R, Hashim S. Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. Applied Psychological Measurement. 2003;27(5):372–79.

25. Rosseel Y Iavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 2012;48:1–36.

26. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Med Care. 2006;44(11 Suppl 3):S152–70. [PubMed: 17060822]

27. Crane PK, Cetin K, Cook KF, Johnson K, Deyo R, Amtmann D. Differential item functioning impact in a modified version of the Roland–Morris Disability Questionnaire. Qual Life Res. 2007;16(6):981–90. [PubMed: 17443419]

•28. Hays RD, Calderón JL, Spritzer KL, Reise SP, Paz SH. Differential item functioning by language on the PROMIS® physical functioning items for children and adolescents. Qual Life Res. 2018;27(1):235–47. [PubMed: 28875367] Hays and colleagues demonstrate methods for examining the impact of differential item functioning

••29. Verdam MG, Oort FJ, Sprangers MA. Item bias detection in the Hospital Anxiety and Depression Scale using structural equation modeling: comparison with other item bias detection methods. Qual Life Res. 2017;26(6):1439–50. [PubMed: 27943018] Verdam and colleagues present a cohesive discussion of extensions to the binary test item, two group, unidimensional latent trait conditions for conceptualizing and evaluating measurement bias.

30. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;2(8):e124. [PubMed: 16060722]

31. Yang FM, Heslin KC, Mehta KM, Yang C-W, Ocepek-Welikson K, Kleinman M, et al. A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. Psychological test and assessment modeling. 2011;53(4):440–60. [PubMed: 23471423]

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

33. Thissen D IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Chapel Hill: University of North Carolina, LL Thurstone Psychometric Laboratory 2001.

34. Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous-DFIT framework. Applied Psychological Measurement. 1999;23(4):309–26.

35. Crane P, Gibbons L, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Med Care. 2006;44(11 Suppl 3):S115–S23. [PubMed: 17060818]

36. Muraki E, Bock D. PARSCALE for Windows. Scientific Software International, Chicago, 2003.

37. Muthén L, Muthén B. Mplus Users Guide. Eighth ed Muthén & Muthén: Los Angeles, CA; 1998–2017.

38. Finch W The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. Applied Psychological Measurement. 2005;29(4):278–95.

39. Finch W, French BF. Detection of crossing differential item functioning: A comparison of four methods. Educational and Psychological Measurement. 2007;67(4):565–82.

40. Finch W, French B. Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. Educational and Psychological Measurement. 2008;68:742–59.

41. French BF, Maller SJ. Iterative purification and effect size use with logistic regression for differential item functioning detection. Educational and Psychological Measurement. 2007;67(3): 373.

42. Stark S, Chernyshenko OS, Drasgow F. Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. J Appl Psychol. 2006;91(6): 1292–306. [PubMed: 17100485]

43. Zwick R, Thayer DT, Wingersky M. A simulation study of methods for assessing differential item functioning in computerized adaptive tests. Applied Psychological Measurement. 1994;18(2):121–40.

44. Wang W-C. Assessment of differential item functioning. Journal of Applied Measurement. 2008;9(4):387–408. [PubMed: 19092232]

45. Woods CM, Grimm KJ. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. Applied Psychological Measurement. 2011;35(5):339–61.

46. Muthén B Beyond SEM: General latent variable modeling. Behaviormetrika. 2002;29(1):81–117.

47. Jones R, Gallo J. Education and sex differences in the mini-mental state examination: Effects of differential item functioning. Journals of Gerontology Series B-Psychological Sciences and Social Sciences. 2002;57(6):P548–P58.

48. Fratiglioni L, Jorm AF, Grut M, Viitanen M, Holmen K, Ahlbom A, et al. Predicting dementia from the Mini-Mental State Examination in an elderly population: the role of education. J Clin Epidemiol. 1993;46(3):281–7. [PubMed: 8455053]

49. Wu X, Sawatzky R, Hopman W, Mayo N, Sajobi TT, Liu J, et al. Latent variable mixture models to test for differential item functioning: a population-based analysis. Health and quality of life outcomes. 2017;15(1):102. [PubMed: 28506313]

50. Peng R, Dominici F, Zeger SL. Reproducible Epidemiologic Research. Am J Epidemiol. 2006;163(9):783–89. [PubMed: 16510544]

51. Rothman KJ, Greenland S, Lash T. Modern Epidemiology. third ed Wolters Kluwer, Lippincott Williams & Wilkins; 2008.

52. Wiegand RE. Performance of using multiple stepwise algorithms for variable selection. Stat Med. 2010;29(15):1647–59. [PubMed: 20552568]

53. Chun S, Stark S, Kim ES, Chernyshenko OS. MIMIC methods for detecting DIF among multiple groups: exploring a new sequential-free baseline procedure. Applied psychological measurement. 2016;40(7):486–99. [PubMed: 29881065]

Box A.

Commonly Used Synonyms for Differential Item Functioning

Factorial invariance
Item bias, Item response bias
Invariance, non-invariance
Factorial invariance
Measurement bias
Measurement (non-)invariance
Measurement confounding
Metric invariance
Scalar invariance
Systematic measurement error

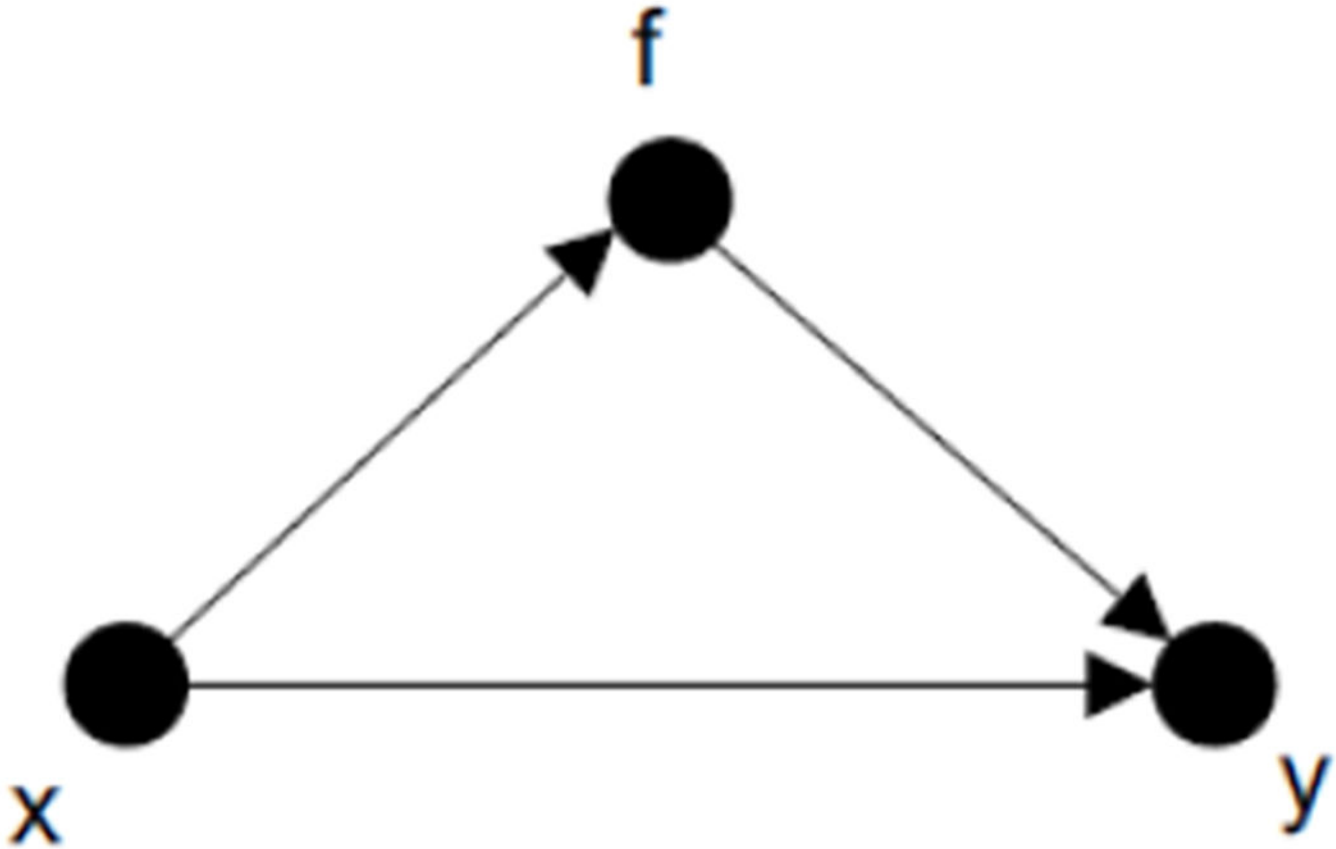

Figure 1. DAG illustrating a basic kind of DIF.

This figure illustrates a basic kind of DIF, where a covariate or predictor (*x*) has a direct relationship with an item (*y*) even after controlling for how the covariate or predictor relates to the underlying construct (*f*).

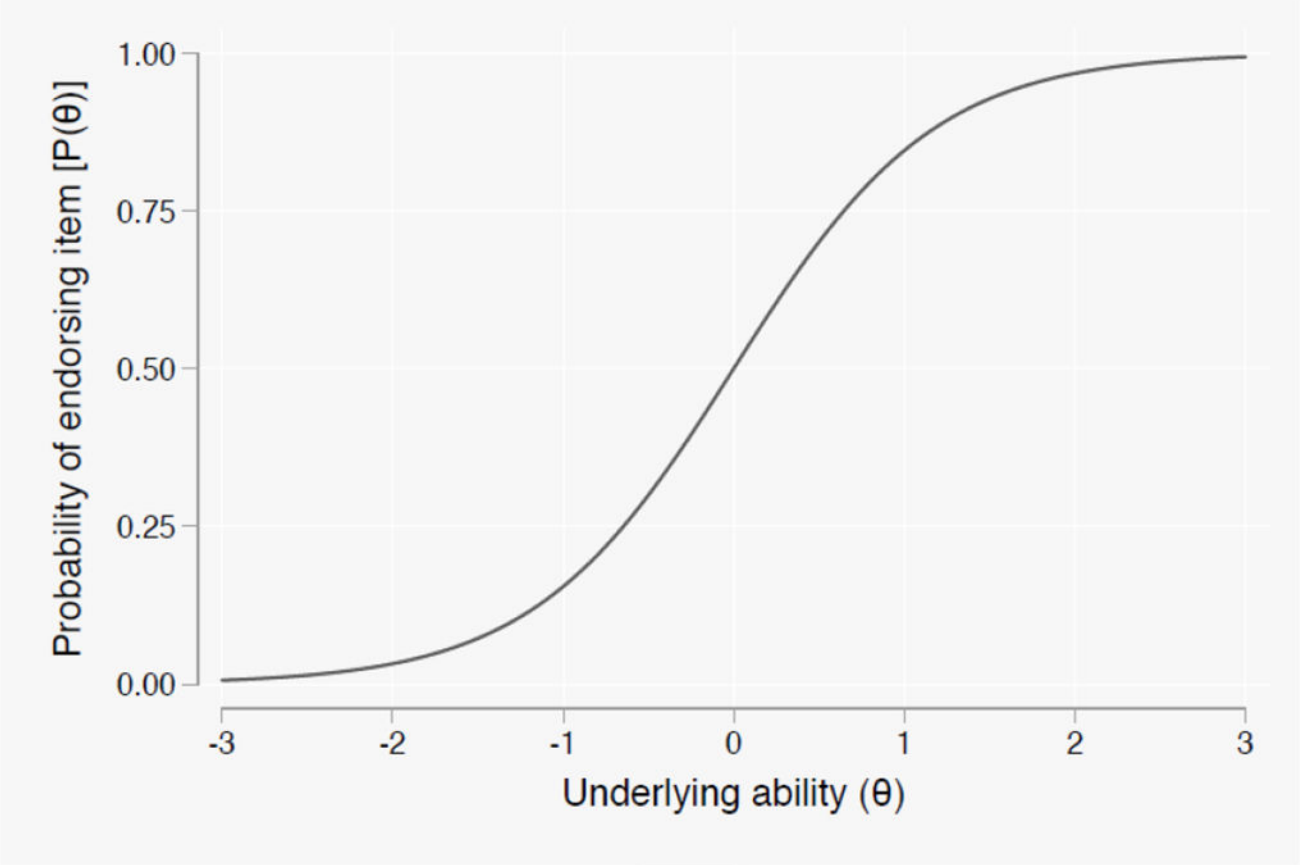

Figure 2. Item characteristic curve

Figure 2 displays a hypothetical item characteristic curve (ICC) for a binary test item. The y-axis indicates the probability of a randomly selected person from the population endorsing the symptom or responding correctly to the test item. The x-axis indicates the level of the underlying or latent trait presumed to cause the item response. The scale of the x-axis reflects the conventional normal distribution assumption for the latent trait.

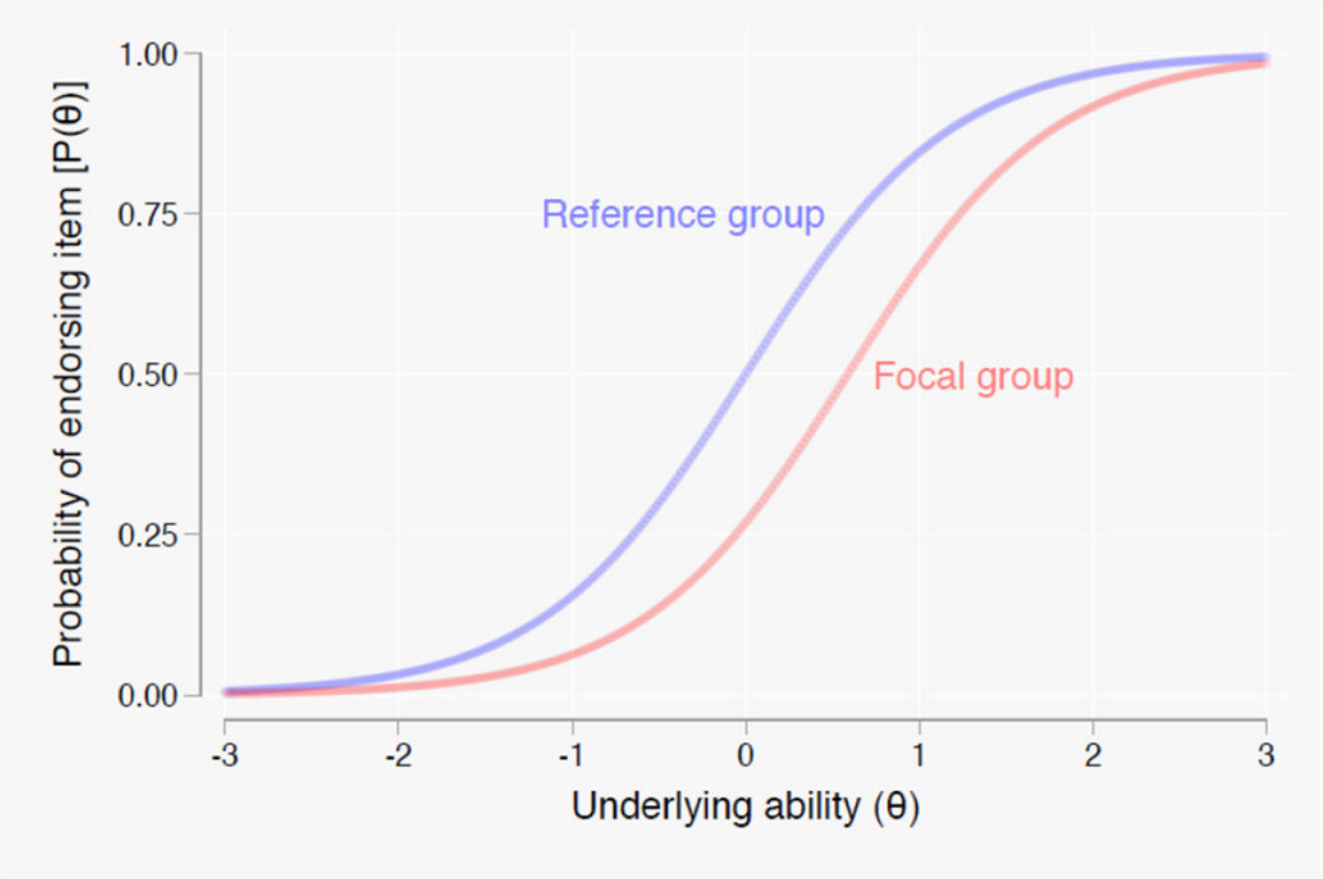

Figure 3. Item characteristic curves for a single binary item in two groups illustrating uniform differential item functioning

Figure 3 displays a hypothetical item characteristic curves for two groups. Within a non-linear transformation, the two lines are parallel. This is indicative of uniform differential item functioning.

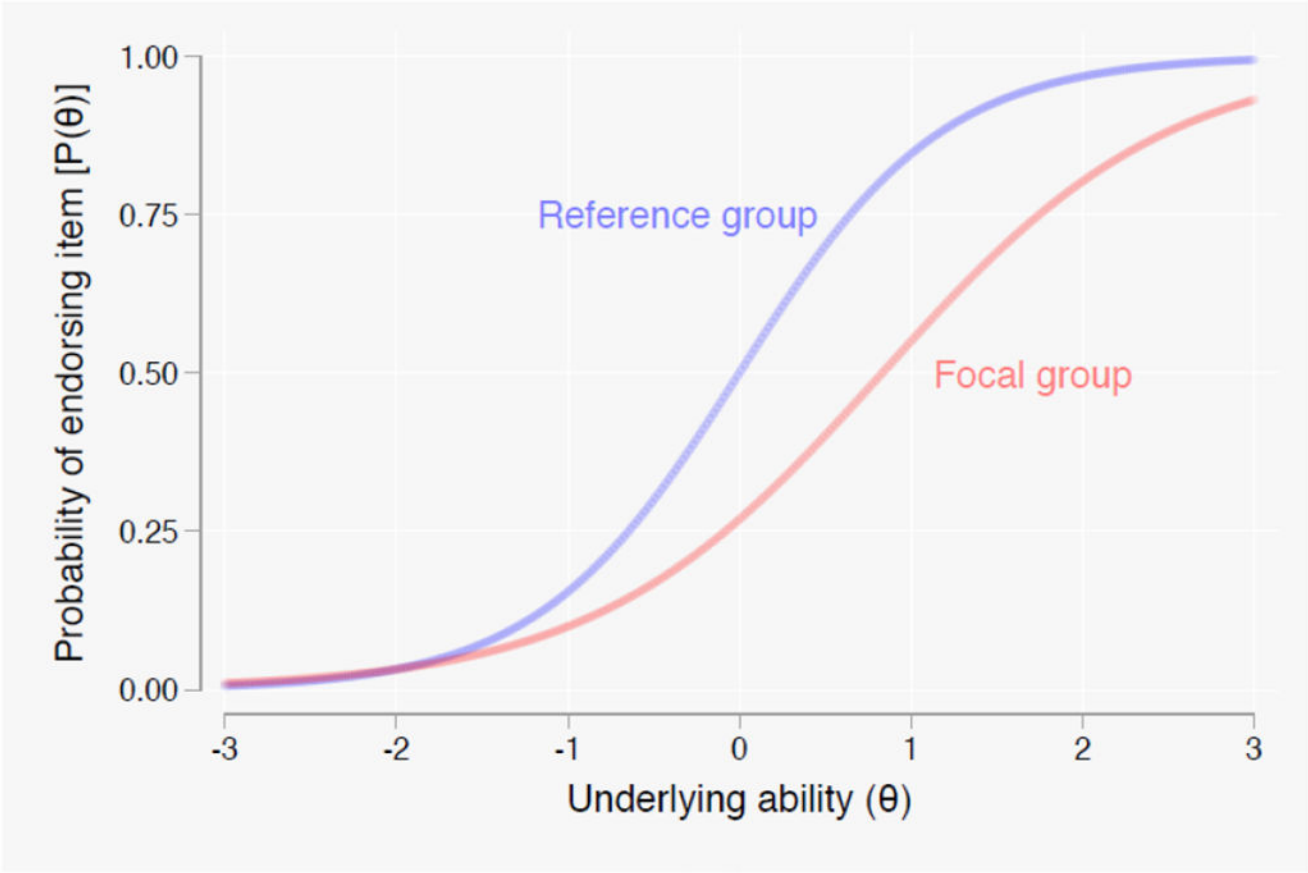

Figure 4. Item characteristic curves for a single binary item in two groups illustrating non-uniform differential item functioning

Figure 4 displays a hypothetical item characteristic curves for two groups. The two ICCs are not parallel, even within a non-linear transformation. This is indicative of non-uniform differential item functioning.

Table 1.

Features and considerations of DIF detection alaorithms

Anchor items	Anchor items are items that are identified prior to the main DIF analysis and are constrained to have no DIF in the DIF detection procedure. These could be items identified before data analysis that -- on the basis of theory, substantive knowledge, prior research, or empirically based on preliminary analysis -- the data analyst assumes will not demonstrate DIF in the analysis. *A priori* chosen anchors also helps protect against the problem of constant bias. Constant bias refers to the situation where all items in the test function differentially and in a similar fashion for members of the focal group relative to the reference group [7]. DIF detection methods will fail to find any items with DIF in such situations, in the absence of declared anchors.
Purification	When the DIF detection algorithm involves a stepwise procedures and separate DIF detection and IRT calibration (estimation of the latent trait value), it may be appropriate to update the IRT ability estimate on the basis of modified item parameters given items identified with DIF continually throughout the detection algorithm.
Multiple comparisons correction	Depending on the algorithm, DIF detection may involve testing multiple hypotheses. As such, it seems reasonable to consider using significance thresholds that account for this multiplicity. Some algorithms have been described that use Bonferroni correction, Benjamini-Hochberg false discovery procedures, and no correction. There is little guidance in the field supported by simulation studies as to which or how multiple comparisons correction should be used. The best answer for one procedure might not be the best answer for another algorithm.
Forward vs backward identification of items with DIF	Most DIF detection algorithms proceed in a forward stepwise fashion. That means, initial models assume all items are invariant across group and obtain test statistics or other misfit indicators for each item in the test (excluding anchors, as appropriate). Given the well-known challenges with stepwise selection in regression models, it is surprising that there is not a preference for backwards selection [52,53].
Covariates	Some methods for DIF detection allow for the inclusion of covariates [26]. In the context of educational and achievement testing, it is plausible that the absence of covariate handling is to be expected due to homogeneity of tested samples (e.g., high school seniors taking a college admissions test). In the context of epidemiology, where heterogeneity of tested samples might be expected to be greater, the absence of covariate handling can be a critical issue for consideration of an appropriate method. In the absence of covariate handling other strategies such as matching and weighting could be used to control for confounding.
Flagging rules	Some methods of DIF detection will identify items with DIF on the basis of statistical significance, change in model fit, or effect size of the detected DIF, or even a combination of these factors. Insofar as the need and character of multiple comparisons correction is not clearly understood, a standard nomenclature and benchmarks for effect size have not been adopted, and different estimation methods offer different approaches to model fit, this is an important area for the emergence of discrepancies across approaches.