Papeles del Psicólogo, 2010. Vol. 31(1), pp. 75-84
http://www.cop.es/papeles

BIAS IN MEASUREMENT INSTRUMENTS. FAIR TESTS

Juana Gómez-Benito*, M. Dolores Hidalgo and Georgina Guilera***
**Universityof Barcelona and **University of Murcia*

Psychological assessment must ensure the equity and validity of interpretations and of any decisions taken as a result of them. Therefore, it is necessary to use bias-free assessment instruments capable of evaluating the personal and social needs of individuals with different characteristics. The study of possible bias in tests, or in some of their items, has been of great relevance in psychometric research over the last 30 years, and is likely to remain an important focus of interest for professionals and researchers involved in psychological and educational testing. The aim of this paper is to provide the applied psychologist with some background and guidelines in relation to bias, the concepts of differential functioning and impact, procedures for detecting item or test bias and the evaluation of its possible causes, with a view to improving the validity of psychological measurement.
Key words: *Bias, Differential Item Functioning (DIF), Detection procedures, Fair tests, Validity.*

Las evaluaciones psicológicas deben garantizar la equidad y validez de las interpretaciones y decisiones adoptadas a partir de las mismas. Para ello es necesaria la utilización de instrumentos libres de sesgo, y capaces de evaluar necesidades personales y sociales de individuos con diferentes características. El estudio sobre el posible sesgo de los tests, o de parte de sus ítems, ha ocupado un lugar relevante en la investigación psicométrica de los últimos 30 años y es previsible que siga constituyendo un importante foco de interés para los profesionales e investigadores implicados en la evaluación mediante el uso de los tests. Este trabajo pretende abordar esta perspectiva ofreciendo al psicólogo aplicado unas directrices y un bagaje de conocimientos sobre los conceptos de sesgo, funcionamiento diferencial e impacto, los procedimientos de detección de ítems o tests sesgados y la evaluación de sus posibles causas para, en conjunto, mejorar la validez de las mediciones psicológicas.
Palabras clave: *Sesgo, Funcionamiento diferencial del ítem, Procedimientos de detección, Tests justos, Validez.*

Tests constitute one of the standardized measurement instrumentsmost widely used in the social and health sciences, especially in psychology and education. It should be borne in mind that a test is administered with a specific objective, generally to make decisions which in most cases are relevant to the life of the examinee. Thus, for example, in Spain tests are used in the recruitment of security guards and of employees in general, in the university context for assessing students, to assess participants in intervention programmes, and so on. It is therefore of extreme importance for the professionals who employ these types of instrument to guarantee equality of opportunity and equal treatment for those to whom they are administered; in other words, to ensure that tests and the decisions made on their basis are fair.

But when can we state that a test is fair?Deciding to what extent a test is fair with respect to its measurement is not an easy task. Aspects such as the sociocultural context, the process ofits construction and/or adaptation, the conditions of application, the interpretationof the scores and the extent of training of the professional administering the test (Muñiz&Hambleton, 1996) can result in the instrument being unfairly employed. As Muñiz and Hambleton themselves point out, the majority of the problems involved with tests derive from their inappropriate use, more than from the test itself, its construction or itstechnical properties. Therefore, assuming that the first two questions have been dealt with, the focus turns to the technical or psychometric properties of the test.

Correspondence: Juana Gómez-Benito. *Dpto. Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Paseo Valle Hebrón, 171. 08035-Barcelona. España. E-mail: juanagomez@ub.edu*

BIAS, IMPACT AND DIF

In this context, the presence of possible bias in the items making up a test is ofprime concern in evaluating the validityofmeasurement instruments, validity being understood as the extent to which the empirical evidence and theoretical reasoning support the appropriateness of the interpretations based on the scores in accordance with the proposed uses of the test (Messick, 1989; Prieto&Delgado, 2010). Thus, when we state that a particular test is valid, what we are actually saying is that

the score obtained has a specific meaning, assuming that this meaning is the same in the different groups for which the test has been validated. Nevertheless, in order to guarantee that a score on a test has the same meaning in different groups, it is necessary to carry out numerous studies that evaluate different evidence of the test's validity(APA, AERA,&NCME, 1999). The existence of bias in psychologicalmeasurement instrumentscan represent a serious threat to the validity of those instruments in which some of the items are benefiting certain groups of the population to the detriment of others with the same level in the trait to be measured. Likewise, an absence of bias in the items constitutes evidence of the degree of generalizationof the interpretations based on the test scores for the different subgroups of one or several populations.

The issue of bias has been of considerable concern for researchers and professionals, especially in the wake of the controversy generated by Jensen's (1969, 1980) studies. This author proposed that intelligence was hereditary, and hence, that the differences observed between racial groups were attributable to genetics. Such a claim obviously sparked lively discussions between the "nature" and "nurture" schools of thought. According to the latter, the explanation for the differences between the groups was to be found in the potential cultural bias of intelligence tests. At that moment, the role of psychometriciansinvolved ascertaining the extent to which the differencesbetween groups were due to real characteristics of the individuals in each group or to artifactsgenerated by the instrument itself. This debate gave rise to a new semantic conflict: cultural bias or different psychometric properties?

Bias refers to the injustice deriving from one or various items of the test on comparing different groups that occurs as a consequence of some characteristicof the item or of the test's application contextwhichis irrelevant for the attribute measured by the item; differences of psychometric properties, on the other hand, refer only to the item's statistical characteristics. Today there is a consensus as regards the term bias, whereby it is assumed that the causes of certain items behaving differentially as a function of certain variables are either known or under study; however, in the majority of studies, all that can be inferred is that there are differences in the item responses obtained by examinees of equal ability. The appropriate term for this latter type of results, which refer only to psychometric properties, is *Differential Item Functioning* (DIF), after the work of Holland and Thayer (1988), who set out to distinguish between the two concepts.

Formally, it is stated that a given item presentsDIF if at apsychometric level it behaves differentially for different groups. In other words, DIF indicates a difference in the functioning of the item (or test) between comparable groups of examinees, being understood as comparable those groups that have been matched with respect to the construct or trait measured by the test (Potenza&Dorans, 1995). Thus, an item presents DIF when groups of equal ability show a different probability of answering it correctly or in a given direction depending on the group to which they belong.In DIF terminology, the name *focal group* is given to the set of individuals, generally a minority, that represents the study's focus of interest and which is normally the disadvantaged group, whilst the term *reference group*refers to a set of standard individuals, generally a majority, with which the focal group is compared. Nevertheless, the fact that a measurement instrument produces systematically poorerresultsin one group compared to another does not necessarily imply the presence of DIF, since there could be real differences between the groups in the trait measured by the test in question. In such cases we talk about *impact* (Camilli&Shepard, 1994) or valid differences (van de Vijver&Leung, 1997).

Having clarified the difference betweenbias, DIF and impact,let us imagine we are studying an item potentially biased against a minority group.How can we assess the presence of DIF? Logic would probably lead us to compare directly the scores on the item obtained by the minority group and by the rest of the examinees, and if we found differences, to say that the item is unfair to one of the groups. However, we cannot be certain whether the differences derive from the item bias or whether the ability levels of one group and the other are actually different. The conceptof DIF sets out to deal with this question, so that DIF analysis compares item responses between groups only when the groups have been matched in level of the measured ability or trait by means of a matching criterion. Thus, it is essential to have a bias-free criterion; nevertheless, in the majority of situations the onlyavailable empirical evidence about the ability level of a examinee is the test itself (generally the total score), which is contaminated by the presence of items with DIF and which form part of the criterion together with the DIF-free items. Therefore, a problem endemic to DIF detection methods lies in the fact that their procedures are somewhat circular,

since the item studied also contributes to the definition of the matching variable for the groups. For reducing the effect ofitemswithdifferential functioning, some purification techniques have been proposed which, in two stages or iteratively, remove from thecriterion those itemspreviously detected as presenting DIF (French&Maller, 2007, Gómez-Benito&Navas, 1996; Hidalgo&Gómez-Benito, 2003; Holland&Thayer, 1988; Navas-Ara&Gómez-Benito, 2002; Wang, Shih, &Yang, 2009).

TYPES OF DIF

Although there are various DIF taxonomies (see Hessen, 2003), one of the most widely used classifications, not least because of its simplicity, is that proposed by Mellenbergh (1982). This author distinguishes two types of DIF according to whether or not there is interactionbetween the level in the measured attribute and the group to which the individuals belong. In so-called *uniform* DIF there is no interaction between level of the measured trait and membership of a particular group, since the probability of respondingto the item correctly (or in a particular direction) is greater for one group than for the otherin uniform fashion across all the levels of the trait. In the case of *non-uniform* DIF there is such interaction, and the probability of each group responding correctly (or in a particular direction) to the item is not the same for all the different levels of the measured trait.

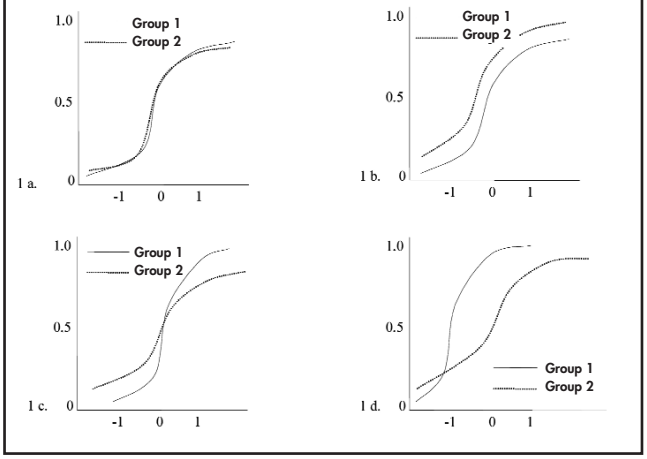
Within the framework of item response theory(IRT) (seeMuñiz[2010] in this issue) the concept of *Item Characteristic Curve* (ICC) is proposed, of great utility for understanding in graphical fashion the diverse types of DIF. In dichotomous response items, ICC relates the probabilityof a correct response to the item (the y-axis on the graph) with the individuals' level in the measured variable or ability (x-axis). Thus, an item does not present DIF if itscharacteristic curvefor thefocal group and for thereference group overlap(Figure 1a), a situationthat occurs when both the difficulty parameter (position of the ICC on the ability scale) and the discrimination parameter (proportional to the slope of the ICC) show a similar value in each group. The item shows *uniform* DIF if the respective ICCs do not cross at any level of the measured variable(Figure 1b), which occurs when the difficulty parameters are different, but the corresponding discriminationparameters remain equal in the two groups. Finally, *non-uniform* DIFis presented if the ICCs cross at some point. In this last case, Swaminathanand Rogers (1990) establish a second subdivision. *Symmetrical non-uniform* DIFwould be represented by a central crossing of the ICCs at the ability level(Figure 1c), and occurs when the difficulty parameterremains constant and the discrimination parametervaries between the two groups, whilst *mixed non-uniform* DIFis found when the difficulty and discrimination parameters are different in the two groups, and is represented by an asymmetric crossing of the ICCs of the focal and reference groups (Figure 1d).

DETECTION PROCEDURES

From the late 1980s and throughout the nineties, the development and analysis of statistical methods and techniques for DIF detectionand evaluationwere the focus of research efforts, as a result of which the procedures employed gradually became more sophisticated. The principal methodological challenge was to developprocedureswhich, on the one hand, are sensitive enough to detect both uniform and non-uniform DIF, and on the other, do not confuse DIF with impact. Furthermore, a growing demand for techniques applicable to polytomous items (such as those which use Likert-type scales) gave rise to the development of procedures that were useful for this type of response format, generally deriving from extensions of their counterparts for dichotomous items.

Taking into account this initial distinction about the nature of the item response (dichotomous/polytomous), Potenza andDorans (1995) classify the different methods according to the type of criterion used for matching the groups (observed score/latent variable) and to the

FIGURE 1
ICC OF AN ITEM WITHOUT DIF (1a), WITH UNIFORM DIF (1b), WITH SYMMETRICALNON-UNIFORM DIF (1c) AND WITH MIXED NON-UNIFORM DIF (1d)



relationship between the score on the item and the matching variable (parametric/non-parametric). Based on this taxonomy, Hidalgo and Gómez-Benito (2010) offer a classification of all current proceduresfor the detectionof DIF.

First of all, the ability level of the individuals can be estimated following two strategies: the first, the *latent variable method*,uses an estimationof the latent ability in the framework of item response theory, whilst the *observed score method* consists in using the observed total scoreon the test. A second criterion concerns how the item score at each ability level is estimated. One way of proceeding is to use a mathematical function that relates the item score with the ability level, such as the ICCs inFigure 1, which represent graphically the probability of obtaining a given score on the item according to the individuals' ability level. As already mentioned, differencesin the ICCs of the groups indicate DIF, and for this to occur, the parametersthat define the corresponding ICCs must be different. Given that the curves are determined by one or more parametersin the mathematical function, this approach is referred to as the *parametric method*. On the other hand, the second strategy does not use any mathematical function to relate item response and ability level; rather, it simply takes into account the observed item score at each ability level for each group. In this case, the presence of DIF will be determined by the obtaining of differences between groupsin the observed score, without taking into account any mathematical model (or, therefore, any parameters). For this reason this approach is known as the *non-parametric method*. Thirdly, the nature of the response type, dichotomous orpolytomous, is considered. Given that in the case of polytomous items DIF can be present in the different response categories of the same itemand not necessarily in the same direction or in all the categories, the techniques fordichotomous itemsare always computationally and conceptually simpler than the extensions for polytomous response items.

Techniques that employ the observed score on the test as a matching variable, assuming that this score is an adequate estimator of the individual's latent ability, can turn out to be inaccurate in the detection of DIF mainly when the test contains items that vary in discrimination. However, the latent variable methods overcome this shortcoming through the increased sophistication of the mathematical models for estimating ability. An advantage of non-parametric methods, such as Mantel-Haenszel(MH) and SIBTEST, is that the model's assumptions are weak, so that DIF is not usually confused with a lack of fit of the model. In the case of parametric methods, such as the procedures based on IRT, it is necessary to guarantee adequate estimation of item parameters in order to avoid such confusion, so that much larger sample sizes of the reference and focal groups are required than withnon-parametric models.

There is a wide range of computer programs that permit the implementation of the majority of DIF detection procedures. Most are programs designed specifically for the detection of items with DIF, such as MHDIF (Fidalgo, 1994), EZDIF (Waller, 1998a), DIFAS (Penfield, 2005) or EASYDIF (González, Padilla, Hidalgo, Gómez-Benito, &Benítez, 2009) for the MH procedure, and available free of charge from the program's authors; DIF/DBF (Stout&Roussos, 1999) for the SIBTEST procedure, distributed through the Assessment System Corporation; RLDIF (Gómez-Benito, Hidalgo, Padilla,&González, 2005) for the Logistic Regression (LR) procedure, and which is about to come onto the market; and IRTLRDIF (Thissen, 2001), TESTGRAPH (Ramsay, 2000) and LINKDIF (Waller, 1998b) for procedures based on IRT, also freely distributed. It is also possible to use resources from standard statistical analysis programs, which require a licence for their use, such as SPSS (SPSS Inc., 2009) for MH, andLR, LISREL (Jöreskog&Sörbom, 2006) or MPLUS (Muthén&Muthén, 1998, 2007) for procedures based on structural equation models.

A wide variety of studies have explored DIF detection techniques using data simulation design in both dichotomous and polytomous item. These studies basically analyze the variationin the rate of false positives or Type I error (detecting an itemwith DIF when in reality it has none) and of correct detections or statistical power (identifying an item with DIF when it actually presents it) under different simulation conditions, manipulating those variables that can supposedly modulate the corresponding detectionrates (e.g., sample size, test contamination or type of DIF) and observing the changes that occur in them.Such studies generally conclude by make suggestions and recommendationsabout the conditionsunder which the procedure in question presents control of Type I error rate and adequatestatistical power. A feature common to practically all of these studies is that they focus on the detectionof DIF in a singleitem. It should be borne in mind that a test is obviously made up of a set of items, and that

the direction of the DIF in the various items of the same test may be different (some may favour the focal group and others the reference group), so that the individual effects of the DIF of the items cancel each other out on considering the test as a whole. Therefore, it is sometimes relevant to evaluate the so-called *Differential Test Functioning* (DTF) or to explore DIF in a subset ofitems. In this context, some techniques have set out to deal specifically with the study of DIF in tests or sets of items, such as SIBTEST in dichotomous itemsand POLYSIBTEST in polytomousitems, or to approach the question from IRT, as proposed byRajuand his research team (Oshima, Raju,&Nanda, 2006).

EFFECT SIZE

Another type of study, also based on data simulation, advises the inclusion of effect size measuresas a complementor alternative tosignificance tests, with a view to increasing the magnitude of the effect observed and comparing the results obtained in differentstudies. It should be borne in mind that detecting an item with DIF through a test ofstatistical significancedoes not necessarily imply that its effect is notable; indeed, the effect may be of small relevance. In this regard, it is important to examine the magnitude of the DIF because the effects of the presence of itemswith DIF can be trivial, can cancel each other out or can actually call into question the decisions based on the test. The majority of DIF-detection techniques have proposed diverse measures. By way of example, Dorans and Holland (1993) present the Delta-DIF statistic for the MH procedure, and in the framework of LR, Zumbo and Thomas (1997) have suggested the increase in R^2; Gómez-Benito and Hidalgo (2007) and Monahan, McHorney, Stump and Perkins (2007) have proposed using the odds-ratio as a measure of effect size,employing LR for dichotomous items, and Hidalgo, Gómez-Benito and Zumbo (2008) proposed the same measure for polytomousitems. As a general rule, these works set guidelines or propose classification criteria that permitthe interpretation of the values of DIF magnitude (not just the presence or absence of DIF), in line with the classification guidelines of the *Educational Testing Service*, which establish threecategories: insignificantDIF (category A), moderateDIF (category B) and high DIF (category C). Items classified as type C should be reviewed and removed from the test, while those classified as types A and/or B can remain in the test.

CHOICE OF TECHNIQUE

Despite considerable progress in the development and optimizationof DIF detectionmethods, the suitability of applying a given procedure in a specific situation still raises many questions. Within this network of research findings and techniques, the doubts usually boil down to the following question: which procedures do we use with our data? The decision to apply one technique or another tends to be based on diverse considerations, since there is so far no method appropriate for all situations. The variables usually taken into account includedifferencesin the ability distributions of the reference and focal groups, the sample size of both groups, the type of DIF, the computational simplicity and availability of the computer programs, and the matching criteria for the groups, among others. And this complexity has led various authors to the conclusion that the most conservative option is to apply variousDIF detection techniques and make the final decision to maintain, reformulate or remove the item according to the convergence or divergence between detection methods, taking into account the characteristics and peculiarities of each procedure. It seems evident that if various techniques coincide in their decisions, there is more certainty about the presence or absence of DIF, whilst if there is divergencebetweentechniqueswe should focus our attention on the characteristicsof the detection proceduresemployed. In any event it would be a case of accumulating evidence in one direction or the other, as in any procedurefor the validation of an instrument.

Following the classificationof detectionmethodsbased on the type of criterion for matching the groups (observed score/latent variable) and the relationship between the scoreon the item and the matching variable(parametric/non-parametric), we can identify four types of detection methodsfor both dichotomous and polytomous response items: i)observed score/parametric, ii) observed score/non-parametric, iii) latent variable/parametric, and iv) latent variable/non-parametric. As mentioned above, methods which useobserved scoreas an estimationof the individuals' ability can beinaccuratewhen thematchingcriterion presents ahigh percentageofitemsthat functiondifferentially, whilst latent variablemethods canovercome this failing by increasing the mathematical complexity.But anadvantage ofnon-parametricmethodsis that the assumptionsof the model are weak, so that DIF is not usually confused with lack of fit of the model, while with parametric methods it is necessary to ensure adequate model fit so as to avoid

suchconfusion, and hence much largersample sizes are required than in the case of non-parametric models. Bearing in mind the general pros and cons of the different types of detection techniques, it might be recommended to make the final decision based on the applicationof one technique of each of the four existing types. For example, one option would be to use LR, MH, IRT and SIBTEST in the detection ofdichotomous items. However, it would be necessary to take into account other aspects of the data that could explain possible divergences betweendetection methods.

The first of these aspects is related to sample size. If one works with small samples, it has been seen that LR and MH function adequately fordichotomous items(Muñiz, Hambleton,&Xing, 2001; Swaminathan&Rogers, 1990) andIRTdoes so (using the likelihood-ratio test)forpolytomous items (Bolt, 2002). With more considerable sizes one can opt for othertechniques, such as SIBTEST and POLYSIBTEST for the detectionofdichotomous and polytomousitems.

A second variable to be considered is the type of DIF. Some techniqueshave been designed specifically for the detection of uniform DIF, so that they may present certain difficulties for the detection of non-uniform DIF,whilst others have been proposed for the detectionofboth types of DIF. When the presence of non-uniform DIF is suspected it is preferable to use techniquesthat are sensitive to this type of differential functioning. Once again, with small sample sizes one can opt forLR withdichotomous items (Hidalgo&López-Pina,2004) andIRT (using the likelihood-ratio test)with polytomous items (Bolt, 2002). For larger sizes one might choose other techniques, such as SIBTEST in the case ofdichotomous itemsandmultinomial logistic regression(Zumbo, 1999) orDFIT (Oshima, Raju, &Nanda, 2006) in that of polytomousitems.

It is also clear that the majorityof methodscurrently used for detecting DIF require that the test to be analyzed contains alarge number of items (e.g., more than 30) for the result to be reliable. But the questionnaires and surveys customarily used in the social and health sciencestend to havesmall numbers of items (between 5 and 30 items). On working with such short tests thereliabilityof the scores is lower, so that the measurement errors are greater.The DIF-detection effectiveness of methods such as LR or MH, which use the observed scoreon the test as a matching variable in the DIF analysis, can be seriously affected. The use of MIMIC models (Gelin&Zumbo, 2007) is an alternative.

Given that most studies postulate that with the application ofproceduresfor purifying the criterionthere is a reduction of the false positives rate and an increase in the statisticalpower of variousmethods,the use of such purification procedures is advisable. Finally, as far as possible it is recommended to accompany detection rates with some measure of effect size.

FAIR TESTS

We have already mentioned how in the 1970s researchers in the USA began to question the use of tests for assessing different groups of individuals in an equitable way, and how the article by Jensen (1969) on the hereditary nature of intelligenceintensified the debate between the nature and nurture camps. This controversyhad considerable social andpoliticalrepercussions, to the extent that tests indicating differences according to socioeconomic or racial characteristics were considered biased and unfair. Indeed, courts began to pass sentences invalidating decisions on employee recruitment or admission to educational institutions. One of the most relevant consequences was the so-called "golden rule", which emerged from the agreement between the Educational Testing Service (themostprominent test company in the USA) and the Golden Rule insurance company, and according to which those items on which white examinees'scores were 15% higher than those of black examinees were to be removed. Evidently, this rule, based only on the item difficulty index for different groups, could result in the removal of itemswith high discriminative power for the measured trait.

At that time the terms bias and injustice were considered equivalent, and there were no effective criteria for identifying whether thedifferential functioning of the test was due to actual differencesin the trait or to artifactualdifferences deriving from the instrumentemployed.The opposition to the use of tests for making decisions affecting the employment and academic contexts also served as an incentive for psychometriciansto make greater efforts to provide definitions andtechniquesfor the detectionof bias, giving rise to one of the most fruitful lines of psychometric researchof recent decades. Thus, in the 1970s and 80s we see the emergence of theterm "differential item functioning" as distinct from "bias", the differencesbetween DIF and impact are highlighted and detection techniques are proposed that permit the

differentiation of the two aspects.With the 1990s comes the explanation of DIF by means of the dimensionality of tests; thus, Ackerman (1992) distinguishes between objective ability (that which the test sets out to measure) and noise ability (which is not intended to be measured but which may influence the responses to some test items): DIF can be presented if the items in the test measure a noise ability in which respondents differ according to their group. Roussos and Stout (1996) refine the issue, changing the terminology to speak ofsecondaryabilitiesrather than noise ability and distinguishing between benign DIF and adverse DIF. Benign DIF occurs when the secondary ability is an auxiliary dimension which the testers want to measure, and adverse DIF is found when the secondary ability is a noise ability.

In any case, and while it is crucial to providestatisticalprocedurescapable of effectively detectingitems withdifferential functioning, these do not in themselves offer an explanation of why the DIF occursand whether or not it implies bias. It should not be forgotten that the presence of DIF is a necessary but not a sufficient conditionfor being able to talk about item bias: DIF exists when individuals of comparable abilitybut from different groups responddifferentially to the item, while for bias to exist it is also necessary for those differencesto be attributable to somecharacteristicof the itemthat has nothing to do withthe attribute measured by the test.

Toward the end of the 1990s and the beginning of the new century, researchers began to stress the importance of analyzing the causes ofDIF. In the contextof test adaptation, the studiesby Allalouf, Hambleton and Sireci (1999) and by Gierl and Khaliq (2001) point to some possible causes revolving around item format and content; Zumbo andGelin (2005) also recommend the consideration of diversecontextual variables. However, Ferne and Rupp (2007), in a review of 27 studiesthat attempt to identify the causes of DIF, argue that the progress made is of scarce relevance. This may be one of the great challenges for DIF researchtoday, and which merits the same determination in research efforts as the problems which, as we saw earlier, have already been solved.To this end it would be advantageous to carry out studiesexpresslydesigned for exploring the causes ofDIF, and multidimensional theorycan undoubtedly orient the search for causes toward those spurious abilitiesdifferentially distributed among the groups compared. Considering a result withDIF as evidence ofbias involves explainingwhy the trait is multidimensional for a specificsubgroupand setting out a reasoned argument for the irrelevanceof thesource of DIF for that trait (Camilli&Shepard, 1994).

Finally, as with any other aspect of validity, the analysisof DIF is a process involving the accumulation of evidence. Assessing and interpreting such evidence requires the rational judgement of experts, and there is no single correct answer. In this regard, we must appeal to the professionalresponsibility of those who use tests and increase our awareness with regard to the relevance of the metric quality of these instruments, with the ultimate aim of guaranteeing a fair and appropriate measurement process. As a preliminary step in the application of detection methods, Hambleton and Rogers (1995) drew up a list of indicators that may arouse suspicion of the presence of DIF – items that associate men with sport and women with the home, that use certain words whose meaning is more familiar for one culture than for another (food, games, illnesses, historical events, etc.), and so on. Moreover, Hambleton (2006) recommends both the developers and the users of tests to take into account previous research on DIF, which can provide information about the common characteristics of items with DIF, as well as on the peculiarities shared by items without differential functioning. Such information is crucial both for the development of newitemsand for alerting us to the possible presence of DIF in existing tests.

As Zumbo (2007) notes, methods for the detection of DIF and item bias are typically used in the process of item analysis on developing newmeasurement instruments, on modifying existing tests for a new assessment context or populations not considered in the instrument's original design, on adapting tests to other languages and cultures, or on validating the inferences drawn from test scores. Clearly, then, the ambit of application of DIF analysis is broad, and covers the various phases of the design and adaptationof a measurement instrument. InSpain, where the majority of tests are imported, the analysis of DIF and bias is of particular importance in the adaptation of standardized instrumentsto our own language and cultural context. Under the auspices of the Spanish Psychological Association (*ColegioOficialde Psicólogos*), a set of guidelines has been drawn up specifically for use in the creation and adaptation of tests, and in which DIF obviously has a prominent role (Muñiz&Hambleton,1996).

No less relevant is the role of DIF in the latest *Standards for educational and psychological testing*(APA et al., 1999), which include DIF and bias analysisin the consideration of validity,and more specifically in relation to evidence based on the internal structure of the test.In sum, thedecisionon whether or not the result obtained in a study is evidence of biascan only be taken based on validity theory: knowing the theory underlying the test, the interpretation to be made of the scores, and the context in which the test is used; in this regard, the broadening of the content of validity permits studieson bias to approach the social perspective of the problem as one more facet of the process of validationof a test. The article by Prieto and Delgado (2010) in this same issue describesthe validationprocess in more detail.

FURTHER INFORMATION

Readers interested in more in-depth knowledge about DIF detection techniquesand the practical implicationsof the presence of DIF in test items might wish to consult the various theoretical reviews (Camilli&Shepard, 1994; Fidalgo, 1996; Gómez-Benito & Hidalgo, 1997; Hidalgo&Gómez-Benito,1999, 2010; Osterlind&Everson, 2009; Penfield&Lam,2000; Potenza&Dorans, 1995), which approach the study of the different techniquesin narrative fashion, outlining the procedures, highlighting the advantages and disadvantages of their application, and making some recommendationsfor their use.

ACKNOWLEDGEMENTS

This work was funded in part by the Spanish Ministry of Science and Innovation (PSI2009-07280) and also by the Government of Catalonia (Generalitat de Catalunya) (2009SGR00822).The authors would like to thank VicentePonsodafor his invitation to participate in this special issue and the rest of the contributors for helping to get this project off the ground.

REFERENCES

Ackerman, T.A. (1992). A didactic explanation of items bias, item impact, and item validity from a multidimensional perspective.*Journal of Educational Measurement, 29*, 67-91.

Allalouf, A., Hambleton, R. K. &Sireci, S. G. (1999).Identifying the causes of DIF in translated verbal items.*Journal of Educational Measurement, 36*(3), 185-198.

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

Bolt, D.M. (2002).A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods.*Applied Measurement in Education, 15*(2), 113-141.

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

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

Ferne, T. & Rupp, A. A. (2007).A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Fidalgo, A.M. (1994). MHDIF – A computer-program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18*(3), 300-300.

Fidalgo, A. M. (1996). Funcionamientodiferencial de los ítems [Differential item functioning]. In J. Muñiz (Coord.), *Psicometría* (pp. 370-455), Madrid: Universitas.

French, B.F.&Maller, S.J. (2007).Iterative purification and effect size use with Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement, 67*, 373-393.

Gelin, M.N. &Zumbo, B.D. (2007).Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales.*Journal of Modern Applied Statistical Methods, 6*, 573-588.

Gierl, M. J., &Khaliq, S. N. (2001).Identifying sources of differential item and bundle functioning on translated achievement tests.*Journal of Educational Measurement, 38*, 164-187.

Gómez-Benito, J., & Hidalgo, M.D. (1997).Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica [Evaluation of differential functioning in dichotomous items: A methodological review]. *Anuario de Psicología, 74*(3), 3-32.

Gómez-Benito, J. & Hidalgo, M.D. (2007). Comparación de varios índices del tamaño del efecto en regresión

logística: Una aplicación en la detección del DIF [Comparison of various effect size indices in logistic regression: An application in the detection of DIF]. Communication presented at: *X Congreso de Metodología de las Ciencias Sociales y de la Salud, Barcelona, 6-9 febrero.*

Gómez-Benito, J., Hidalgo, M. D., Padilla, J. L., & González, A. (2005). Desarrollo informático para la utilización de la regresión logística como técnica de detección del DIF [Computer development for the use of logistic regression as a DIF detection technique]. Computer demonstration presented at: *IX Congreso de Metodología de las Ciencias Sociales y de la Salud, Granada, Spain.*

Gómez-Benito, J., & Navas, M.J. (1996). Detección del funcionamiento diferencial del ítem: Purificación paso a paso de la habilidad [DIF detection: Step-by-step purification of the skill]. *Psicológica, 17*, 397-411.

González, A., Padilla, J.L, Hidalgo, M.D., Gómez-Benito, J. &Benítez, I. (2009) EASY-DIF: Software for analysing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement.*(Submitted for publication).

Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11 Suppl. 3), S182-S188.

Hambleton, R.K., & Rogers, H.J. (1995). Item bias review (EDO-TM-95-9). Washington, DC: Clearinghouse on Assessment and Evaluation.

Hessen, D.J. (2003). *Differential item functioning: Types of DIF and observed score based detection methods.* Dissertation (supervisors: G.J. Mellenbergh& K. Sijtsma). Amsterdam: University of Amsterdam.

Hidalgo, M. D., & Gómez-Benito, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos [Techniques for the detection of DIF in polytomous items]. *Metodología de las Ciencias del Comportamiento, 1*(1), 39-60.

Hidalgo, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*(1), 1-11.

Hidalgo, M. D., & Gómez-Benito, J. (2010).Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw(Eds.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier - Science & Technology.

Hidalgo, M.D., Gómez-Benito, J. &Zumbo, B.D. (2008).Efficacy of R-square and Odds-Ratio effect size using Discriminant Logistic Regression for detecting DIF in polytomous items. Paper presented at the *6th Conference of the International Test Commission, 14-16 July, Liverpool, UK.*

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(4), 903-915.

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

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*(1), 1-123.

Jensen, A. R. (1980). *Bias in mental testing.*New York: Free Press.

Jöreskog, K.G., & Sörbom, D. (2006).*Lisrel 8 (version 8.8).* Chicago, Illinois: Scientific Software International, Inc.

Mellenbergh, G. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Messick, S. (1989).Validity.In R. Linn (Ed.).*Educational measurement* (3rd edition, pp. 13-104). Washington, DC: American Council on Education.

Monahan, P.O., McHorney, C.A., Stump, T.E. & Perkins, A.J. (2007).Odds-ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression.*Journal of Behavioral Statistics, 32*,1, 92-109.

Muñiz, J. (2010). Las teorías de los tests: TeoríaClásica y Teoría de Respuesta a los Ítems [Test theories: Classical Theory and Item Response Theory]. *Papeles del Psicólogo, 31*(1), 57-66.

Muñiz, J., & Hambleton, R.K. (1996). Directrices para la traduccion y adaptacion de los tests [Guidelines for test translation and adaptation]. *Papeles del Psicólogo, 66.*

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

Muthén, L.K., &Muthén, B.O. (1998, 2007). *MPLUS statistical analysis with latent variables.User's Guide.* Los Angeles, CA: Muthén and Muthén.

Navas-Ara, M. J. & Gómez-Benito, J. (2002). Effects of

ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*(1), 9-15.

Oshima, T. C, Raju, N. S. & Nanda, A. O. (2006).A new method for assessing the statistical significance in the differential functioning of item and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Osterlind, S. J., & Everson, H. T. (2009).*Differential item functioning* (2nd edition). Thousand Oaks, California: Sage Publications, Inc.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*(2), 150-151.

Penfield, R. D., & Lam, T. C. M. (2000).Assessing differential item functioning in performance assessment: Review and recommendations.*Educational Measurement: Issues and Practice, 19*, 5-15.

Potenza, M., &Dorans, N. (1995).DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Prieto, G. & Delgado, A. (2010).Fiabilidad y validez [Reliability and validity].*PapelesdelPsicólogo, 31*(1), 67-74.

Ramsay, J. O. (2000). *TestGraph: A program for the graphical analysis of multiple choice and test questionnaire*. Unpublished manual.

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

SPSS 15.0. (2009). SPSS Inc. 1989-2009.

Stout, W. & Roussos, L. (1999).*Dimensionality-based DIF/DBF package* [Computer Program]. William Stout Institute for Measurement.University of Illinois.

Swaminathan, H. & Rogers, H. J. (1990).Detecting differential item functioning using logistic regression procedures.*Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D. (2001). *IRTLRDIF v2.0b.Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Test for Differential Item Functioning*.Available on Dave Thissen's web page.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage Publications.

Waller, N. G. (1998a). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and Logistic Regression procedures. *Applied Psychological Measurement, 22*, 391.

Waller, N.G. (1998b). LINKDIF: Linking item parameters and calculating IRT measures of Differential Item Functioning of Items and Tests. *Applied Psychological Measurement, 22*, 392.

Wang, W.-C., Shih, C.-L.& Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning.*Educational and Psychological Measurement, 69*, 713-731.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D.&Gelin, M. N. (2005).A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5*, 1-23.

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