Psicothema

Psicothema 2018, Vol. 30, No. 1, 104-109
doi: 10.7334/psicothema2017.183

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Differential Item Functioning: Beyond validity evidence based on internal structure

Juana Gómez-Benito[1], Stephen Sireci[2], José-Luis Padilla[3], M. Dolores Hidalgo[4], and Isabel Benítez[5]
[1] Universidad de Barcelona, [2] University of Massachusetts Amherst (USA), [3] Universidad de Granada, [4] Universidad de Murcia, and Universidad Loyola Andalucía[5]

Abstract

Background: In the latest release of the Standards for Educational and Psychological Testing, Differential Item Functioning (DIF) is considered as validity evidence based on internal structure. However, there are no indications of how to design a DIF study as a validation study. In this paper, we propose relating DIF to all sources of validity evidence, and provide a general conceptual framework for transforming "typical" DIF studies into validation studies. **Method:** We perform a comprehensive review of the literature and make theoretical and practical proposals. **Results:** The article provides arguments in favour of addressing DIF detection and interpretation as validation studies, and suggestions for conducting DIF validation studies. **Discussion:** The combination of quantitative and qualitative data within a mixed methods research perspective, along with planning DIF studies as validation studies, can help improve the validity of test score interpretations.

Keywords: DIF, validity, sources of validity evidence.

Resumen

Funcionamiento Diferencial del Item: más allá de las evidencias de validez basadas en la estructura interna. **Antecedentes:** en la última edición de los Standards for Educational and Psychological Testing, el Funcionamiento Diferencial del Item (DIF) es considerado como una evidencia de validez basada en la estructura interna. Sin embargo, no hay indicaciones claras sobre cómo diseñar un estudio de DIF como un estudio de validación. Proponemos relacionar el DIF con todas las fuentes de evidencias de validez y un esquema conceptual para transformar los estudios "típicos" de DIF en estudios de validación. **Método:** se lleva a cabo una extensa revisión de la literatura y realizan propuestas teóricas y prácticas. **Resultados:** el artículo aporta argumentos a favor de abordar la detección e interpretación del DIF como estudios de validación y recomendaciones para realizar estudios de validación sobre el DIF. **Discusión:** la combinación de resultados cuantitativos y cualitativos en un marco de investigación mixta, junto con el diseño de los estudios de DIF como estudios de validación, puede mejorar la validez de las interpretaciones de las puntuaciones en los tests.

Palabras clave: DIF, validez, fuentes de evidencias de validez.

Test theory has come a long way since, at the beginning of the twentieth century, Spearman (1904) outlined the first ideas concerning Classical Test Theory. Since then, numerous advances have been made that have changed the way in which the social and educational sciences have approached measurement. There have also been changes in the way in which main measurement concepts such as "validity" have been understood, particularly in relation to test and item bias. This evolution can be traced by following the series of releases of the *Standards for Educational and Psychological Testing*, published by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council of Measurement in Education (NCME), from the first version published in 1954 through the sixth edition published in 2014.

Received: May 18, 2017 • Accepted: November 30, 2017
Corresponding author: Juana Gómez-Benito
Facultad de Psicología
Universidad de Barcelona
08035 Barcelona (Spain)
e-mail: juanagomez@ub.edu

The study of bias in items and tests began at the end of the 1960s and developed exponentially in the following decades, in part due to its deep social, psychological and educational implications. In the introduction to the 1974 *Standards*, the authors declared "… part of the stimulus for revision is an awakened concern about problems like… or discrimination against member of groups such as minorities and women" (APA, AERA, & NCME, 1974, p. 1). Social justice in the form of interest in the equal treatment of ethnic and socio-economical groups has been a determining factor in stimulating the study of the item and test bias.

The bias of measuring instruments has emerged as something more than a purely technical issue in psychometric analysis; it has become a subject of educational, social, and legal debate. For example, the *Golden Rule* case (*Golden Rule Life Insurance Co. et al. v. Mathias et al., 1980*) led to the development of methods for identifying *Differential Item Functioning* (DIF) to screen out items on employment tests that might be biased against particular subgroups of examinees. In the late nineties, two special issues in *Educational Measurement: Issues and Practices*, addressed the heated debate between advocates and critics of considering testing consequences as a validity issue (Crocker, 1996). Currently,

there is renewed attention to equity and fairness in assessment including a broader conceptualization of validity evidence needed to justify the use of a test for a particular purpose (Sireci, 2016), or, for instance, the debate about the degree to which large-scale educational assessments have accomplished their intended goals of improving instruction and educator effectiveness (Lane, 2014). This growing interest can also be noticed outside the United States. The latest version of the model proposed by the European Federation of Psychological Association to assess the quality of tests includes DIF as one of the possible research designs to gather evidence of construct validity (Evers et al., 2013). In Spain, the evaluation of psychological and educational tests carried out by the *Test Commission of the Spanish Psychological Association* also paid attention to DIF and fairness issues (Hernández, Tomás, Ferreres, & Lloret, 2015; Prieto & Muñiz, 2000).

The latest edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) state "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Thus, any test or item parameter that is different between two or more subpopulation groups, like item difficulty or discrimination, may be a threat to validity if the difference would require different interpretations for each group. In other words, measurement by means of tests may be invalidated by the presence of items that show different psychometric properties across groups of people from different demographic, social, cultural, or linguistic backgrounds. In this context, our thesis is DIF becomes an important validity issue for test score interpretations. The rationale behind our theoretical proposal comes from the concept of "construct representation" (Embretson, 1983), which includes as threat to validity "construct-irrelevant variance". DIF analyses can become part of a comprehensive validation effort aimed to identify what parts of the assessment are reliable variance associated with other constructs, response sets, guessing, etc., all of them irrelevant to the intended construct (Messick, 1989).

The AERA et al. (2014) *Standards* describe five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). The *Standards* describe DIF analyses only as validity evidence based on internal structure. At the same time, the *Standards* assign a key role to DIF when addressing fairness in testing issues. Although we agree that DIF studies are important to deal with fairness issues, unlike the *Standards*, we do not confine DIF results to validity evidence based on internal structure. We see DIF as more than a "dimensionality problem" because interpretation of DIF results is necessary for improving our measurements and improving validity of test score interpretations. As the AERA et al. (2014) *Standards* describe,

"However, Differential Item Functioning is not always a flaw or weakness. Subsets of items that have a specific characteristic in common (e. g., specific content, task representation) may function differently for different groups of similarly scoring test takers. This indicates a kind of multidimensionality that may be unexpected or may conform to the test framework" (p. 16).

We view DIF evidence as a broader issue than validity evidence based on internal structure. Going beyond the source of validity evidence based on internal structure, allows researchers to delve into whether DIF items are capturing something different across groups and how DIF results threaten the intended interpretation of test scores. Thus, the aim of this paper is twofold: (1) to highlight the central role of DIF studies on the validity of test score interpretations in a more comprehensive way, so that is not confined to the evidence related to the internal structure, but rather to all five sources of validity evidence outlined in the *Standards*; and (2) to demonstrate how DIF studies can be addressed as validation studies within the *Standards* as the validation framework.

Evolution of concepts and methods… bringing DIF closer to validity!

The metric characteristics of assessments should be invariant across different groups on whose performance on the test is intended to make *valid* comparative interpretations. In this context, an item exhibits DIF when its psychometric properties differ across groups after the groups have been matched on the trait or ability measured by the test (Angoff, 1993). However, conceptualizations of how these properties may differ have expanded over time. For example, Mellenbergh (1982) defined non-uniform DIF in dichotomous items and Penfield and Lam (2000) introduced a taxonomy of "differential step functioning" to interpret DIF in polytomous items (pervasive vs. non pervasive DIF; constant, convergent and divergent DIF, plus the combinations). This growing "arsenal" of terms can be seen as researchers' efforts to characterize and better understand DIF effects and their causes.

Since the beginning of DIF and bias research, one of the core problems has been to figure out when differences between groups are artifacts caused by the measurement process itself and, therefore, outside the intended use given to the test, or are real. Group differences in item performance that represent a difference in the construct measured are traditionally referred to as *impact*, representing a construct-relevant difference (Camilli & Shepard, 1994). Avoiding the confounding of "impact" and DIF has been, and is still, a permanent concern in item bias research.

DIF analyses seek to flag items for *potential* bias by identifying items on which differential group performance, beyond that expected by true group differences, is observed. From our perspective, distinguishing DIF from impact, and determining whether DIF items are measuring the intended construct, are fundamental validity issues in pursuit of fairness in testing.

Given that DIF analyses are only a preliminary step in the evaluation of item bias, there has long been interest in distinguishing these two terms from one another. The term "Differential Item Functioning" appeared in the literature after the term "item bias" to emphasize the statistical nature of DIF (Holland & Thayer, 1988). DIF involves only a statistical analysis, while item bias involves the combination of a statistical finding with a substantive explanation regarding the construct-under representation and/or construct-irrelevant cause of the differential item performance.

Keeping a clear distinction between DIF and item bias is becoming increasingly difficult as statistical DIF methods are more sophisticated, and new contexts for DIF studies appear beyond traditional monolingual comparative groups formed by demographic variables (Gómez-Benito, Balluerka, González, Widaman, & Padilla, 2017). Sireci (2005a) pointed out how common assumptions underlying traditional DIF methods became less tenable when DIF is extended to cross-lingual comparisons (e.g., difficulties in analyzing DIF without considering translation issues, contextual differences between groups, etc.). What is more, DIF methods are also used to validate test accommodations for special populations (Sireci, 2005b).

A look at the evolution of DIF detection methods illustrates a permanent concern about the extent to which DIF results represent threats to the validity of test score interpretations. A comprehensive review of the techniques and new application of DIF techniques would be departing from the scope of this work (see Hidalgo & Gómez-Benito, 2010). In this section, we only focus on methods that have historical significance in the tension between DIF and item bias as a validation issue.

The first DIF detection proposals come on the heels of Cleary and Hilton (1968) who used analysis of variance (ANOVA), and Angoff and Ford (1973) who proposed the delta-plot analysis. Both techniques adjust only for the overall mean differences across groups and so they were called "unconditional" in the sense that it does not match the groups across all levels of the trait measured. On the other hand, "conditional" DIF detection methods try to deal with mean differences across groups and the election of *valid* matching criteria. The Mantel-Haenszel statistic (Holland & Thayer, 1988) has been considered the gold standard in evaluating DIF, and logistic regression (Swaminathan & Rogers, 1990) has gained popularity because it provides a flexible framework for analyzing uniform and nonuniform DIF (Hidalgo, López-Martínez, Gómez-Benito, & Guilera, 2016).

Other procedures that historically played a key role are those that integrated DIF and bias to help interpret the causes of DIF. Cohen and Bolt (2002) noted that the usual strategy for assessing DIF is not ideal for understanding its causes. The traditional approach to DIF is "exploratory," which defines the characteristics of interest of the subjects, but not the dimension that causes DIF. In DIF research it is usual to work with observed variables (sex, language, culture) that we associate with the differential behavior of the item; and when we find differences in item performance in terms of these variables, it is often not apparent why it occurs.

From the perspective of Cohen and Bolt (2002), assessment of the causes of DIF may be more successful if we consider the presence of latent classes in the data. The analysis strategy would be: 1) identify groups of examinees for which the differential item functioning is greater, 2) investigate the characteristics of examinees ranked in these latent groups and determine if the DIF is associated with certain observed characteristics of the examinees assessed. Samuelson (2005) pointed out that among the limitations of the traditional paradigm of DIF is the assumption of intragroup homogeneity (e.g., students with disabilities are sufficiently homogeneous that they can be considered a single group). Zumbo (2007) could include these studies as examples of the third generation of DIF studies, referring to those which conceives DIF as a result of the item characteristics and/or "testing situation factors" that are not relevant to the intended construct. This view is key for our proposal. A step further in this direction is the proposal of the so-called "ecological models" of item responding and DIF (Zumbo, Li, Wu, Shear, Olvera, & Ark, 2015). These models expand latent class techniques for detecting DIF to allow researchers to take factors like personal variables and context into account beside item and tests characteristic to explain DIF.

DIF and the Standards

Given the discussion in the previous sections, and since our aim is to extend DIF to all sources of validity evidence, it is of interest to perform a more detailed analysis of how DIF and its relation with validity are addressed in the latest edition of *Standards*. As mentioned earlier, the AERA et al. (2014) *Standards* refer to DIF as an example of validity evidence based on internal structure. As they describe,

"Some studies of the internal structure of tests are designed to show whether particular items may function differently for identifiable subgroups of test takers (e. g., racial/ethnic or gender subgroups). *Differential item functioning* occurs when different groups of examinees with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item" (p. 16).

However, the term DIF or the concept itself is not present in *any* of the 25 validity standards associated with "Evidence regarding internal structure" in the cluster for "Specific forms of validity evidence." There is just one implicit reference to DIF issues, without labeling it as such, in the comment of standard 1.25 on evidence based on *consequences of tests*:

"When unintended consequences appear to stem, at least in part, from the use of one or more tests, it is special important to check that theses consequences do not arise from construct irrelevant components or construct underrepresentation. For example, although group differences, in and of themselves, do not call into question the validity of a proposed interpretation, they may increase the salience of rival hypotheses that should be evaluated as part of the validation effort" (p. 30).

The rationale behind our thesis that DIF is a validity issue related to all sources of validity evidence borrows largely from the concept of "construct representation" (Embretson, 1983), and considers DIF results as evidence for *rival hypothesis* to the intended test score interpretations for proposed uses of the tests. Therefore, we think that what standard 1.25 states for test consequences can be extended to all sources of validity evidence.

On the other hand, the relation between DIF and testing consequences is addressed in detail in *Fairness in Testing* chapter of the AERA et al. (2014) *Standards*. For example, the chapter includes specific standards for DIF/item bias such as standard 3.6:

"When credible evidence indicates that test scores may *differ in meaning* for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individual from those subgroups" (p. 65).

Following the traditional practice of performing DIF studies during the test development process, mainly as part of item analyses, the *Standards* appeal for DIF in chapter 4 (*Test Design and Development*), stating,

"In theory, the ultimate goal of such studies is to identify construct-irrelevant aspects of item content, item format, or scoring criteria that may differentially affect test scores of one or more groups of test takers. When differential item functioning is detected, test developers try to identify plausible explanations for the differences, and then they may replace or revise items *to promote sound score interpretations for all examinees*" (p. 82, emphasis added).

It is important to note that for the first time in the latest version of the *Standards* (AERA et al., 2014) there is a specific standard addressing the procedural characteristics of DIF analyses related to item development and review. Specifically,

"Statistics used for flagging items that function differently for different groups should be described, including specification of the groups to be analyzed, the criteria for flagging, and the procedures for reviewing and making final decisions about flagged

items. Sample sizes for groups of concern should be adequate for detecting meaningful DIF" (p. 88).

To sum up, DIF is not clearly articulated in the AERA et al. (2014) *Standards* as a validity issue. There are only indirect references like "Where credible evidence indicates that test scores may *differ in meaning* for relevant subgroups…" (p. 65), or the call to test developers "… to promote sound score interpretations for all examinees" (p. 70), that relate DIF to validity issues. Thus, there is a need for a clear articulation of the role of DIF in validating test score interpretations.

General conceptual framework to conduct DIF validation studies

Definitions of validity in the latest editions of *Standards* borrow largely from Messick (1989) who stated "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Before describing the role of DIF analyses in validation of test score interpretations, the arguments of the current consensus about validity are: (a) tests must be evaluated with respect to a particular purpose, (b) what needs to be validated are the inferences derived from test scores, not the test itself, (c) evaluating inferences made from test scores involves several different types of qualitative and quantitative evidence, and (d) evaluating the validity of inferences derived from test scores is not a one-time event; it is a continuous process (e.g., Kane, 2013).

Consistent with the 1999 edition (AERA, APA, & NCME, 1999), the current version of *Standards* introduces validity as a unitary concept and describes five sources of "complementary" validity evidence "that might be used in evaluating a proposed interpretation of test scores for a particular use" (AERA et al., 2014, p. 13). The sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing (Sireci & Padilla, 2014). Like in the 1999 *Standards* they are not distinct type of validity given that validity is a unitary concept. Maybe, trying to overcome the lack of impact of that conception on daily validation practice (Cizek, Rosenberg, & Koons, 2008), the 2014 *Standards* groups some validity standards in cluster 3 "Specific Forms of Validity Evidence". This cluster includes validity standards related to each source of validity evidence.

Two conceptual pillars are needed to develop the general framework which allows relating DIF to all sources of validity evidence and transforming a typical DIF study in a validation study. The first pillar is the argument-based approach to test validation proposed by Michael T. Kane in the last two decades. Kane's approach involves developing an argument to support the use or interpretation of test scores for specific purposes (Kane, 2006). Later, Kane (2013) rectified what he called an "imbalance" between "interpretations" and "uses" giving more weight to "uses," replaced the term "interpretative argument" with "interpretative/use argument" (or IUA), "… where the IUA includes all the claims based on test scores (i. e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (p. 2). "Claims" and "assumptions" in the most recent Kane's view of the argument-based approach to validation are keys to our proposal, because we propose to understand the lack of DIF as "assumptions" that need to be tested to support the interpretative/use arguments.

The second pillar is the "de-constructed" approach to validation (e. g., Sireci, 2016) that uses the *Standards* five sources of validity evidence as a validation framework. This framework involves: a) explicit statement of the purposes of testing; and b) using the five sources of evidence to support those explicit purposes.

Both Kane's and Sireci's "de-constructed" approaches to validation converge to relate DIF to all sources of validity evidence. DIF results can confirm or reject an "assumption" that is supporting an IUA (Kane, 2013), or to respond to a "validity question" entitled by the test purpose statement (Sireci, 2016). Moreover, it is important to note that DIF can relate to *any* of the five sources of validity evidence. In Table 1 we present examples of aims for DIF validation studies for each of five sources of validity evidence.

Acknowledging that DIF analyses are relevant to all sources of validity evidence emphasizes how DIF analyses can be used to promote validity. A "DIF validation study" is a DIF study designed to combine DIF results with other sources of validity evidence. Such combination can allow researchers to connect the source of DIF with sources of construct underrepresentation, construct irrelevant variance, or unintended negative consequences to ensure test scores reflect the same construct for all examinees.

Our theoretical proposal of conceptualizing DIF studies as validation studies can be extended beyond educational testing. As

Table 1
Examples of aims for DIF validation studies

Source of validity evidence	DIF validation studies
Test content	To examine if construct representation is similar for identifiable groups of the intended population
	To examine if there are difference in the accessibility of test content To examine if any content in items flagged for DIF is irrelevant to the construct measured
Response processes	To assess if items test tap the same intended process delineated in the test specification for identifiable groups
Internal structure	To analyze if the relationships among items or part of the test are similar for different groups of test takers (dimensionality) To evaluate whether an item measures a construct-irrelevant dimension for some examinees
Relations to other variables	To analyze if the relationships between item/test responses and additional variable or covariates conceptually related follow the same patterns for identifiable groups of the intended population
Consequences of testing	To examine if unintended consequences of testing arises from construct-irrelevant components or construct underrepresentation (e.g., does eliminating DIF items lead to construct underrepresentation? Does the presence of DIF items lead to different pass rates for identifiable groups?

Kane (2013) pointed out the IUA "… may involve an interpretation in terms of a skill or disposition to behave in some way and allow for a range of possible use" (p. 2).

A mixed methods research framework

Mixed-methods research can be the most appropriate methodological research framework to conduct DIF validation studies. Looking for the benefits of "methodological complementarity" mixed methods studies integrate quantitative and qualitative methods. An introduction to mised-methods research is beyond of the scope of the article (see Creswell, 2015).

There are still few but promising examples of mixed methods studies in test validation. Gadermann et al. (2011) conducted think aloud protocol interviews to examine the cognitive processes of children when responding to scale items, and logistic regression analysis to detect group differences in the cognitive processes. Benítez and Padilla (2014) integrate DIF results and cognitive interviewing findings to interpret DIF. Benítez, Padilla, Hidalgo and Sireci (2016) interpret DIF in PISA 2006 combining DIF quantitative results with expert appraisal contributions to content validity. Maddox, Zumbo, Tay-Lim, and Qu (2015) integrate quantitative DIF results with ethnographic transcript to uncover how Mongolian respondents cope with three items of a literacy test.

Discussion

In this article, we discussed the need for relating DIF analysis to all sources of validity evidence. Despite the consensus reached about DIF concepts and techniques, up to this point DIF analyses have not been properly integrated within a typical test validation framework. As Sireci and Rios (2013) pointed out test developers rarely retain items that display statistically significant and large DIF particularly in large educational assessment projects. On the other hand, test users often interpret test scores without even considering how DIF can affect habitual total-group test score comparisons (Hidalgo, Benítez, Padilla, & Gómez-Benito, 2015). As a consequence, not to conceptualize DIF analyses as validity studies misses the contribution of DIF to promote test score validity and fairness.

We think that DIF analyses can promote validity and that fairness of test score interpretations provided that DIF studies are planned as part of the larger validation effort. That is, DIF analyses should be an integral part of the validity argument. The validity argument should integrate DIF results with quantitative and qualitative validity evidence to properly interpret whether the DIF represents construct-relevant or irrelevant factors, and whether there are any negative consequences associated with the DIF.

Based on our review of the DIF literature, our research practices, and the AERA et al. (2014) *Standards*, we offer the following general recommendations for incorporating DIF analyses into a comprehensive validation effort:

(1) Identify the assumptions supporting the "IUA" or the "validity question" posed by the statement of the test purpose. The assumptions or validity questions should refer to differences in the accessibility of test content, cognitive processes, dimensionality of the test, relations to other variables or covariates, testing consequences, or any combination of validity evidence for identifiable groups of test takers.
(2) Design a mixed methods validation study in which quantitative and qualitative methods for obtaining validity evidence can be integrated to test the assumptions stated in (1).
(3) Conduct DIF analyses following best practices recommendations (Sireci & Ríos, 2013), for selecting the DIF detection methods best suited to the data, using more than one DIF method, effect size measures, replicating DIF results, etc., or,
(4) Resort to resort to the quantitative and/or qualitative method more appropriate for obtaining validity evidence with which DIF results can be interpreted.
(5) Integrate DIF results with other quantitative results and/or qualitative findings to examine DIF assumptions supporting the validity interpretative argument.

Rogers and Swaminathan (2016) point out the combination of cognitive psychology findings and modeling techniques as promising venues to improve our understanding of DIF. DIF study may move beyond item comparability/invariance concerns to focus on invariance at the test score/interpretation level which is the level to make test based decisions. For example, in cross-lingual assessment and in evaluating invariance across platforms (e.g., laptop vs. tablet) we may allow items to show DIF, and use separate parameters for them, but maintain comparability at the total test score level. In any case, we believe our proposal for relating DIF to all five AERA et al. (2014) sources of validity evidence will help promote validity and fairness in educational and psychological assessment.

Acknowledgments

This work was supported by the Spain's Ministry of Economy and Competitiveness [Grant number PSI2015-67984-R), and the Andalusia Regional Government under the Excellent Research Fund [SEJ-6569].

References

American Psychological Association, American Educational Research Association & National Council on Measurement in Education (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

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

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-23). Dublin: Educational Research Center.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*(2), 95-106. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Benítez, I., & Padilla, J. L. (2014). Analysis of non-equivalent assessments across different linguistic groups using a Mixed Methods approach: Understanding the causes of Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research, 8*(1), 52-68. https://doi.org/10.1177/1558689813488245

Benítez, I., Padilla, J. L., Hidalgo, M. D., & Sireci, S. (2015). Using mixed methods to interpret Differential Item Functioning. *Applied Measurement in Education, 29*(1), 1-16. https://doi.org/10.1080/08957347.2015.1102915

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. United States: SAGE Publications.

Cizek, G.J., Rosenberg, S.L., & Koons, H.H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412. https://doi.org/10.1177/0013164407310130

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*(1), 61-75. https://doi.org/10.1177/001316446802800106

Cohen, A.S., & Bolt, D.M. (2002). A mixture model analysis of differential item functioning. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Thousand Oaks, CA: Sage Publications.

Crocker, L. (1996). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*(2), *4*. https://doi.org/10.1111/j.1745-3992.1997.tb00584.x

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197. https://doi.org/10.1037/0033-2909.93.1.179

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283-291. https://doi.org/10.7334/psicothema2013.97

Gadermann, A.M., Guhn, M., & Zumbo, B.D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for children: A focus on cognitive processes. *Social Indicator Research, 100,* 37-60. https://doi.org/10.1007/s11205-010-9603-x

Gómez-Benito, J., Balluerka, N., González, A., Widaman, K. F., & Padilla, J. L. (2017). Detecting differential item functioning in behavioral indicators across parallel forms. *Psicothema, 29*(1), 91-95. https//doi.org/10.7334/psicothema2015.112

Hernández, A., Tomás, I., Ferreres, A., & Lloret, S. (2015). Tercera evaluación de test editados en España [Third evaluation of tests published in Spain]. *Papeles del Psicólogo, 36*(1), 1-8.

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), vol. 4, (pp. 36-44). USA: Elsevier- Science & Technology.

Hidalgo, M. D., Benítez, I., Padilla, J. L., & Gómez-Benito, J. (2017). How polytomous item bias can affect total-group survey score comparisons. *Sociological Methods and Research, 46*(3), 586-603. https://doi.org/10.1177/0049124115605333

Hidalgo, M.D., López-Martínez, M.D., Gómez-Benito, J., & Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema, 28*(1), 83-88. https//doi.org/ 10.7334/psicothema2015.142

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

Kane, M.T. (2006). Validation. In B.L. Robert (Ed.), *Educational measurement* (4th ed., pp. 17-64). Wesport, CT: Praeger

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135. https//doi.org/ 10.7334/psicothema2013.258

Maddox, B., Zumbo, B. D., Tay-Lim, B., & Qu, D. (2015). An anthropologist among the psychometricians: Assessment events, Ethnography, and Differential Item Functioning in the Mongolian Gobi, *International Journal of Testing, 15*(4), 291-309. https://doi.org/10.1080/15305058.2015.1017103

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118. https://doi.org/10.2307/1164960

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: MacMillan.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15. https://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for evaluating quality of test used in Spain]. *Papeles del Psicólogo, 77,* 65-71.

Rogers, J., & Swaminathan, H. (2016). Concepts and methods in research on differential item functioning of tests items. Past, present, and future. In C.S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement* (pp. 126-142). New York: Guilford Press.

Samuelson, K. (2005). Examining Differential Item Functioning from a latent class perspective. *Dissertation Thesis*. University of Maryland.

Sireci, S. G. (2005a). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G. (2005b). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3-12. https://doi.org/10.3102/0013189X034001003

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226-235. https://doi.org/10.1080/0969594X.2015.1072084

Sireci, S. G., & Padilla, J. L. (2014). Validity assessment: Introduction to the special section. *Psicothema, 26*(1), 97-99. https//doi.org/10.7334/psicothema2013.255

Sireci, S. G., & Ríos, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187. https://doi.org/10.1080/13803611.2013.767621

Spearman, C. (1904).The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72-101. https://doi.org/10.2307/1412159

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Zumbo, B.D. (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. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera, O. L., & Tanvinder, K. A. (2015). A methodology for Zumbo's third generation DIF Analysis and the Ecology of Item Responding. *Language Assessment Quarterly, 12*(1), 136-151. https://doi.org/10.1080/15434303.2014.972559