Psicothema

Psicothema 2017, Vol. 29, No. 1, 91-95
doi: 10.7334/psicothema2015.112

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

# Detecting differential item functioning in behavioral indicators across parallel forms

Juana Gómez-Benito[1], Nekane Balluerka[2], Andrés González[3], Keith F. Widaman[4] and José-Luis Padilla[3]
[1] Universidad de Barcelona, [2] Universidad del País Vasco, [3] Universidad de Granada and [4] University of California, Riverside

## Abstract

**Background**: Despite the crucial importance of the notion of parallel forms within Classical Test Theory, the degree of parallelism between two forms of a test cannot be directly verified due to the unobservable nature of true scores. We intend to overcome some of the limitations of traditional approaches to analyzing parallelism by using the Differential Item Functioning framework. **Method**: We change the focus on comparison from total test scores to each of the items developed during test construction. We analyze the performance of a single group of individuals on parallel items designed to measure the same behavioral criterion by several DIF techniques. The proposed approach is illustrated with a dataset of 527 participants that responded to the two parallel forms of the Attention Deficit-Hyperactivity Disorder Scale (Caterino, Gómez-Benito, Balluerka, Amador-Campos, & Stock, 2009). **Results**: 12 of the 18 items (66.6%) show probability values associated with the Mantel $\chi^2$ statistic of less than .01. The standardization procedure shows that half of DIF items favoured Form A and the other half Form B. **Conclusions**: The "differential functioning of behavioral indicators" (DFBI) can provide unique information on parallelism between pairs of items to complement traditional analysis of equivalence between parallel test forms based on total scores.

***Keywords:*** Parallel forms, differential item functioning, differential functioning of behavioral indicators.

## Resumen

***Detección de funcionamiento diferencial del ítem en indicadores conductuales de formas paralelas.*** **Antecedentes**: a pesar de la importancia crucial del concepto de formas paralelas en la Teoría Clásica de los Tests, el grado de paralelismo entre dos formas paralelas no puede comprobarse directamente debido al carácter inobservable de las puntuaciones verdaderas. Nuestra propuesta pretende superar algunas de las limitaciones de los métodos tradicionales utilizando el esquema del Funcionamiento Diferencial del Item. **Método**: cambiamos el objeto de la comparación de las puntuaciones totales a cada uno de los ítems individuales. Analizamos las puntuaciones de un único grupo de participantes en ítems paralelos diseñados para medir los mismos criterios comportamentales. Ejemplificamos la propuesta con las respuestas de 527 participantes a las dos formas paralelas de la "Attention Deficit-Hyperactivity Disorder Scale" (Caterino, Gómez-Benito, Balluerka, Amador-Campos, & Stock, 2009).**Resultados**: 12 de los 18 ítems (66,6%) muestran valores de probabilidad asociados con el estadístico Mantel $\chi^2$ menores de .01. El procedimiento de Estandarización muestra que la mitad de los ítems con DIF favorecen a la Forma A y la otra mitad a la Forma B. **Conclusiones**: el procedimiento "differential functioning of behavioral indicators" (DFBI) puede aportar información única sobre el paralelismo entre parejas de ítems complementando el análisis tradicional de la equivalencia de formas paralelas.

***Keywords:*** formas paralelas, funcionamiento diferencial del ítem, funcionamiento diferencial de los indicadores conductuales.

The measurement of psychological characteristics is extremely complex because such characteristics, or constructs, are latent variables that are therefore not directly observable. As a result, researchers must develop a series of items or other indicators of the target latent variable, and these indicators are then evaluated with regard to their quality in assessing the underlying construct. Procedures followed in test development have been widely studied, and proper approaches to test development boast a long tradition within psychology (Downing & Haladyna, 2006; Martínez, Moreno, Martín, & Trigo, 2009). The current consensus is that the process of test construction must begin with a clear definition of the objectives being sought by the assessment, and test construction should include a series of steps that ensures the quality of the final product. For example, Standard 1.1 included in the most recent published edition of *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & the National Council on Measurement in Education [NCME], 2014) states: "The test developer should set forth clearly how test scores are intended to be interpreted and consequently used." (p. 23).

One of the steps in test construction is the adequate definition of the target variable in terms of the behaviors that are supposed to represent it. Osterlind (1997) recommended that the relations between the behaviors that define the target variable and the

Received: April 29, 2015 • Accepted: October 25, 2016
Corresponding author: Juana Gómez-Benito
Facultad de Psicología
Universidad de Barcelona
08035 Barcelona (Spain)
e-mail: juanagomez@ub.edu

corresponding items be adequately set out in the test specifications. Although the extent to which such information is made explicit may vary across investigations, such specifications should form part of every test development process. Once the behavioral indicators have been specified, more than one item should be developed to measure each behavior that is an exemplar of the construct. In some cases, the abundance of items or other testing requirements may lead to the development of two or more complete forms of a test, which are then referred to as parallel forms. The notion of parallel forms is of crucial importance within classical test theory, as the concept of reliability was first developed in the context of parallel forms (Ferrando, Lorenzo-Seva, & Pallero, 2009). However, the degree of parallelism between two forms of a test cannot be directly verified due to the unobservable nature of true scores. The rationale behind our proposal is to overcome some of the limitations of traditional approach to analyzing parallelism by using the Differential Item Functioning methodological framework.

Parallel models based on less restrictive assumptions than those holding for strictly parallel forms have been proposed - for example, tau-equivalent or congeneric measures (Lord & Novick, 1968; Jöreskog, 1971) - but the definition of these models is still based on hypothesized characteristics of the unobservable true scores, and the degree of equivalence between test forms cannot be directly verified. Therefore, other approaches are usually applied to assess the degree of parallelism of alternate forms indirectly. For example, the Wilks-Votaw-Gulliksen procedure (Gulliksen, 1950) holds that parallel forms should have equal means, equal variances, equal reliabilities, and equal correlations with other variables. Notably, in contrast to procedures used in item response theory, the fundamental level of analysis in classical test theory is the total test score. When studying the degree to which alternate forms of a test meet the above criteria for parallel forms, this focus on the total test scores is maintained and any determination that the alternate forms are strictly parallel or not are based on the total scores of each of the test forms considered as a whole.

The present paper describes a procedure that aims to provide a more accurate assessment of parallelism between test forms during test development by statistical techniques for detecting differential item functioning (DIF). We propose to analyze the degree of equivalence of items designed to measure the same specific behaviour. The proposal leads to a major change in the traditional approach to parallelism: the focus of comparison shifts from qualities of the total test score to qualities of each one of its items. From the DIF perspective, the proposal consists in applying DIF techniques comparing the performance of a single group of individuals on parallel items designed to measure the same behavioral criterion.

DIF methods are currently very sophisticated, and new contexts for DIF studies appear beyond traditional monolingual comparative groups formed by demographic variables. Readers interested in a comprehensive review of DIF methods and practice are referred to the reviews of Hidalgo and Gómez-Benito (2010), or Sireci and Ríos (2013), which also offer a number of guidelines on how to conduct DIF analyses.

As it is well known, in typical DIF studies, one set of items is applied in two populations of individuals, which are referred to as the majority or reference group and the minority or focal group. Then, analyses are performed to evaluate the equivalence of item parameters across the reference and focal groups. However, the way in which DIF detection techniques are used in our methodological proposal varies in two main aspects. Firstly, the procedure described below uses a single group of individuals, rather than two populations of individuals. One key advantage of this approach is that any observed differences or effects across forms at the item level are not attributable to differing characteristics of different samples of individuals answering the sets of items, as only a single sample is used. Secondly, rather than analysing each single item, the proposed procedure analyses each single behavioral criterion for whose measurement more than one item has been developed. Therefore, any differences in performance revealed by the analyses of DIF will be attributable to the specific form or way in which the behavioral indicator was operationalized.

When parallel forms are developed in the terms described above, one would expect that the two items designed to measure a particular behavioral criterion would show equivalent functioning. Thus, if differential functioning is detected, this would indicate that, in some way, the pair of items in question is not actually measuring the same thing in precisely the same way, and the pair of items cannot, therefore, be considered equivalent. The primary goal of the present study is to outline a procedure for evaluating what can be named "the differential functioning of behavioral indicators" (DFBI) that are presumed to be equivalent. The illustrative analysis makes use of a dataset derived from a scale that has two parallel forms and was designed to assess attention deficit hyperactivity disorder (ADHD) in adults.

## Method

### *Participants*

Participants were 527 students (from high schools, junior colleges, and universities) from the southern USA, to whom both forms of the scale were administered in a counterbalanced order. Their mean age was 21.65 years (*SD* = 9.55), and approximately 61% were women.

### *Procedure*

Participants were allowed to complete the scale at their own pace. All respondents took part voluntarily and under conditions of anonymity. Ethical guidelines for research with humans were respected.

### *Instrument*

The scale used to illustrate the procedure for detecting DFBI was the Caterino Adult Attention Deficit-Hyperactivity Disorder Scale (CAADS), whose development and validation process has been described in Caterino, Gómez-Benito, Balluerka, Amador-Campos, and Stock (2009). The scale has two parallel forms (A and B), each comprising 18 items that are indicators of the 18 criteria set out in the *Diagnostic and Statistical Manual of Mental Disorder-Text Revision 4th edition* ([DSM-IV TR] American Psychiatric Association [APA], 2000) to characterize ADHD. Each of the scale items presents respondents with a situation that they must rate on a three-point scale: 0, if it describes their behaviour only a little; 1, if it describes their behaviour to a moderate degree; and 2, if it describes their behaviour to a great degree. Moreover, each of the 18 items has to be answered with regard to three areas:

'at home', 'at work or school', and 'in social settings'. The score for each item is calculated by summing the score obtained in these three contexts, which therefore yields a polytomous item with seven response levels, from 0 to 6. The total score for ADHD is then obtained by summing the score on all 18 items.

*Data analysis*

Data were analysed using the EASY-DIF software package (González, Padilla, Hidalgo, Gómez-Benito, & Benítez, 2011), which enables the user to estimate differential functioning in polytomous items by means of the Mantel statistic (Mantel, 1963) and the standardization procedure (Dorans & Holland, 1993; Zwick, Donoghue, & Grima, 1993), both of which are conditional methods.

Mantel (1963) proposed a statistic that is an extension of the standard Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959), and it is often used as the gold standard for detecting items with differential functioning. It is based on contingency table analysis and consists of comparing the item performance of two groups (reference and focal), which have been previously matched on the ability scale. The matching is done using the observed total test score as a criterion (Holland & Thayer, 1988). In this study, a thick matching strategy (Donoghue & Allen, 1993) was used in order to avoid the excessive numbers of empty cells that would be produced by directly using a thin matching approach such as the total test score. Specifically, the matching strategy was based on collapsing total scores into ability levels until a minimum frequency of 50 individuals per ability level was achieved. The matching criterion used was the total score on the scale and the item analysed was included in the calculation of the final score.

The standardization procedure was first proposed for dichotomous items (Dorans & Kulick, 1986) and later was extended to polytomous items (Dorans & Schmitt, 1991; Zwick et al., 1993). The comparison between reference and focal group is established in relation to the total score in the test grouping in *K* intervals. It is necessary to compute the difference between the mean or the expected value for an item for the focal group and the mean or the expected value for the item for the reference group. A negative value indicates a lower mean for the item in the focal group relative to the reference group, conditioned to the matching variable and DIF favours the reference group. On the other hand, a positive value indicates a higher mean for the item in the focal group than in the reference group and DIF favours the focal group. If the difference is equal to zero, the item is not flawed by DIF.

A confirmatory factor analysis to test the three factor model for CAADS scores were performed by MPlus 7. 3 (Muthén & Muthén, 2014).

## Results

When studying the psychometric properties of the CAADS, Caterino et al. (2009) analysed the equivalence between forms A and B of the test, obtaining a correlation coefficient of .90 between the total scores of the two forms. This value, obtained by applying one of the most widely used methods for examining the degree of equivalence between parallel forms, is indicative of a high degree of equivalence. Moreover, the means and standard deviations of forms A and B were 34.11 (*SD* = 18.64) and 35.62 (*SD* = 19.23), and the coefficient alpha reliability coefficient was .95 for both forms. Regarding the efficiency of the scale in identifying whether individuals met criteria for diagnosis of the disorder, the sensitivity values of forms A and B were .95 and .93, respectively, and their relative specificity values were .86 and .88. Lastly, values of the typical indexes used in confirmatory factor analysis show a good fit of the three factor model for CAADS scores: Root Mean Square Error of Approximation (RMSEA): .050; Comparative Fix Index (CFI): .923; and Tucker Lewis Index (TLI): .911.

All these data provide further evidence regarding the equivalence of the two forms and validity evidence of the internal structure of the scale. However, this approach focuses on total test scores, so is unable to detect the degree of equivalence between each pair of items developed to measure the same behavioral criterion.

Table 1 shows the item statistics and the results obtained when the differential functioning of behavioral indicators or scale items was analysed by means of our proposed procedure, which attempts to overcome the drawbacks associated with methods based on the total scores of different test forms. For each item, the mean, the standard deviation, and its correlation with the scale are provided; the corresponding value of the Mantel $\chi^2$ statistic, its associated probability value, and the value obtained by means of the standardization (STD) procedure are also shown. The results of DIF analysis indicate that a very high proportion of indicators of the same behaviour show statistically significant levels of differential item functioning. Specifically, 12 of the 18 items (66.6%) present probability values associated with the Mantel $\chi^2$ statistic (with one degree of freedom) of less than .01, and two additional items had *p*-values less than .05.

The typical standardization procedure used to detect DIF yields values with a sign. Thus, when used to detect DIF, the typical procedure indicates whether the differential functioning favours or prejudices the minority group. Here, however, only one group of subjects is used, so the sign is associated with one or the other

*Table 1*
Results of the analysis of differential functioning of behavioral indicators on the CAADS

| Item | Form A | | | Form B | | | $\chi^2$ | p | STD |
|---|---|---|---|---|---|---|---|---|---|
| | ME | SD | r | ME | SD | r | | | |
| 1 | 1.572 | 1.769 | 0.610 | 1.841 | 1.962 | 0.497 | 3.047 | .081 | .142 |
| 2 | 2.213 | 2.082 | 0.629 | 2.055 | 2.076 | 0.612 | 8.428 | .004 | -.283 |
| 3 | 1.536 | 1.843 | 0.652 | 1.365 | 1.744 | 0.650 | 12.195 | <.001 | -.273 |
| 4 | 2.302 | 1.878 | 0.522 | 2.536 | 2.011 | 0.531 | 2.291 | .130 | .148 |
| 5 | 1.802 | 1.806 | 0.457 | 2.210 | 2.080 | 0.587 | 12.526 | <.001 | .365 |
| 6 | 1.782 | 1.795 | 0.435 | 1.790 | 1.868 | 0.593 | .629 | .428 | -.052 |
| 7 | 1.430 | 1.761 | 0.548 | 2.133 | 2.035 | 0.585 | 40.905 | <.001 | .622 |
| 8 | 2.588 | 2.111 | 0.624 | 2.497 | 2.116 | 0.542 | 4.540 | .033 | -.217 |
| 9 | 2.195 | 1.882 | 0.545 | 1.923 | 1.956 | 0.603 | 17.199 | <.001 | -.386 |
| 10 | 2.034 | 2.081 | 0.576 | 2.406 | 2.120 | 0.577 | 5.174 | .023 | .224 |
| 11 | 1.206 | 1.772 | 0.606 | 2.174 | 2.103 | 0.608 | 76.468 | <.001 | .834 |
| 12 | 1.721 | 1.970 | 0.598 | 2.350 | 2.086 | 0.448 | 27.778 | <.001 | .508 |
| 13 | 2.155 | 1.865 | 0.340 | 2.226 | 1.953 | 0.405 | .009 | .922 | .022 |
| 14 | 2.191 | 1.960 | 0.432 | 1.476 | 1.898 | 0.510 | 63.977 | <.001 | -.832 |
| 15 | 2.195 | 2.078 | 0.499 | 1.741 | 2.016 | 0.553 | 30.505 | <.001 | -.592 |
| 16 | 1.731 | 1.960 | 0.513 | 2.146 | 1.925 | 0.401 | 10.414 | .001 | .317 |
| 17 | 2.480 | 2.206 | 0.565 | 1.437 | 1.880 | 0.582 | 123.088 | <.001 | -1.184 |
| 18 | 1.121 | 1.615 | 0.557 | 1.593 | 1.861 | 0.531 | 21.571 | <.001 | .403 |

parallel form of the test. The results in Table 1 indicate that, of the items that showed significant differential functioning with respect to the behavioral criterion, half of them favoured Form A and the other half Form B. Purification of the criteria, by eliminating item 17 that shows the greatest magnitude of DIF, did not change the results. All other items with DIF maintained comparable levels of DIF, and one additional item, item 8, now was significant at $p<.01$.

We note that we found a high degree of agreement between the two methods used to detect DFBI. Specifically, the Pearson correlation coefficient between the probability values for the chi-square tests and the absolute value obtained by means of the standardization procedure was very high ($r = .96$), which provides supplementary information that both methods are useful in identifying DIF.

## Discussion

Application of the proposed procedure for detecting DFBI showed the unique information that our new procedure provides. The analysis of equivalence between parallel test forms based on total scores can yield satisfactory results, even though the desired equivalence between pairs of items may not be achieved. For example, the item "I often blurt out answers before questions are completely asked" on Form A of the CAADS and the item "I answer questions before people finish asking them" on Form B are the two alternate versions designed to measure one of the behavioral criteria of impulsivity specified by DSM-IV TR (APA, 2000). When these two items were analysed by means of our proposed procedure for detecting DFBI, the DIF values obtained were $\chi^2(1) = 10.414$ ($p<.001$) and $STD = .317$. In other words, the two alternative wordings that were proposed to be interchangeable in terms of measuring a particular behavioral criterion actually showed differential functioning with respect to this criterion, and the two items cannot therefore be considered equivalent in item functioning.

This information can be very useful for test developers while preparing alternate forms of a test or questionnaire, particularly when they proposed several items for the same construct indicators. DFBI can identify the items that are measuring the construct indicators in an equivalent way when developing parallel or different version of tests or questionnaires. Standard 4. 10 includes a reference to the screening process of item using DIF that could be extended to DFBI: "The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF), (...), should also be documented". (p. 89).

Results shown in Table 1 demonstrate that the proposed procedure for analysing DFBI is useful in terms of identifying situations in which two items designed to measure the same behaviour do not in fact have similar properties. Although previous studies based on total scores obtained with the parallel forms of the CAADS have reported a high degree of equivalence between the two forms, the present results show that a large number of items exhibit differential item functioning. The large number of items with DFBI does not affect the degree of global equivalence between the two forms due to equilibrium in the direction of the differential functioning that arises. That is, approximately as many items favoured Form A as items favouring Form B, so any effects of one set of items compensated for the other set when obtaining the total scores of respondents. This is known as the cancelling-out effect (Nandakumar, 1993) of items that demonstrate DIF in opposing directions. When a cancelling-out effect occurs, analyses of total test scores appear to show an absence of differential functioning across forms. Thus, the procedure proposed here is very useful in the process of test construction as it evaluates the parametric equivalence between all pairs of items that are considered to be equivalent a priori in accordance with the table of test specifications.

The CAADS scale, used here as an illustration, has more than one item for each behavioral criterion, but this is not essential, as our approach to analyses requires only one pair of items to be evaluated for DIF. The remaining common items can be used as anchor items for estimating the ability level of individuals, which is a necessary piece of data in all conditional studies of DIF. For each pair of items whose equivalence is to be evaluated, the process would consist in administering to a single group of individuals the two versions of the item (which would be used as a matching criterion), along with a set of anchor items designed to determine the ability level of these individuals. Of course, the analysis may be based on more than one pair of items or even all items on a test, as was the case with the scale used here, which has two versions or items for each behavioral criterion.

As the proposed procedure is a conditional method, the use of a single group of individuals is not an essential requirement, although it is highly recommended. The advantage of conditional methods resides in eliminating the effect that sample differences across groups could have on results related to DIF testing, because such effects are completely eliminated when analysing data from a single group. In fact, this is identifying unique feature of the proposed procedure, which uses DIF techniques to detect the lack of equivalence between items in a single population of respondents.

Following Ackerman (1992), Kok (1998), and Shealy and Stout (1993), we acknowledge that differential functioning could be due to multidimensionality. However, in contrast to those situations in which DIF techniques are normally used, any multidimensionality cannot represent a secondary dimension that is differentially distributed in the majority and minority groups, because only a single sample is used. In conventional DIF studies, differences on multiple dimensions are found across groups, and items are considered to be constant. By contrast, the way in which DIF techniques were used in the procedure proposed here means that the focal and reference groups are in fact the same group, with the same distributions on all the possible secondary dimensions measured by the items. Therefore, following the current approach, any secondary dimensions can be differentially distributed in the two forms of the item; that is, the so-called parallel items may differ in the additional secondary dimensions measured. In other words, any differential functioning encountered cannot be attributed to the differing characteristics of the multiple samples of individuals, but rather must be due to items that measure specific behaviours differently or, to put it another way, to the lack of equivalence between these items. This specific form of lack of equivalence can be detected by the method proposed here for identifying the differential functioning of behavioral indicators.

Finally, the present study used the Mantel statistic and the standardization procedure, but the logic behind the proposed method means that the most appropriate technique of DIF detection can be applied in each case. The wide range of DIF procedures

available (for recent reviews, see Hidalgo & Gómez-Benito, 2010; Osterlind & Everson, 2009; Sireci & Ríos, 2013) enables their application in a variety of circumstances, for example, with small or large samples, with dichotomous or polytomous items, etc. The proposed procedure is also applicable to test adaptation designs in which a single group of bilingual subjects is used (Sireci, 2005) to analyse the equivalence between two versions of the same test in different languages. In such an application, any differences would again be attributable not to the differential characteristics of individuals, but to the lack of equivalence between the items in the two versions. This highlights the value added by the procedures proposed in this paper.

In summary, when the aim is to examine the equivalence between behavioral indicators of a psychological variable, the proposed method based on DIF detection techniques is more powerful than are classical approaches based on the global correlation between parallel test forms. Indeed, the classical method of estimating the reliability coefficient as a measure of equivalence between different versions of the same test is unable to detect the possible lack of equivalence in some of its items. In this regard, the proposed procedure offers a more exhaustive and precise analysis which, if used during the process of test development, would help to increase the equivalence of parallel forms.

Acknowledgment

This study was partially funded by the Andalusia Regional Government under the Excellent Research Fund (Project nº SEJ-6569, Project nº SEJ-5188), and the Agency for the Management of University and Research Grant of the Government of Catalonia (2014 SGR-1139).

References

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.

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition Text Revision (DSM-IV-TR).* Washington, DC: Author.

Caterino, L., Gómez-Benito, J., Balluerka, N., Amador-Campos, J. A., & Stock, W. A. (2009). Development and validation of a scale to assess the symptoms of Attention-Deficit/Hyperactivity Disorder in young adults. *Psychological Assessment, 21*(2), 152-161.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131-154.

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 Associates.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational Testing Service.

Downing, S. M., & Haladyna, T. M. (2006). Handbook of test development. Mahwah: NJ: Lawrence Erlbaum Associates, Inc.

Ferrando, P. J., Lorenzo-Seva, U., & Pallero, R. (2009). Implementación de procedimientos gráficos y analíticos para la construcción de formas paralelas [Implementing graphical and analytical procedures for developing parallel tests]. *Psicothema*, *21*(2), 321-325.

González, A., Padilla, J. L., Hidalgo, M. D., Gómez-Benito, J., & Benítez, I. (2011). EASY-DIF: Software for analyzing Differential Item Functioning using the Mantel-Haenszel and Standardization procedures. *Applied Psychological Measurement*, *35*, 483-484.

Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale: Lawrence Erlbaum Associates.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric test. *Psychometrika*, *36*, 109-133.

Hidalgo, M. D. & Gómez-Benito, J. (2010). Differential item functioning. In P. Peterson, E. Baker & B. McGaw, (Eds.), *International Encyclopedia of Education*. Volume 4 (pp. 36-44). Oxford: Elsevier.

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.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263-274). New York: Plenum Press.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Martínez, R. J., Moreno, R., Martín, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21,* 326-330.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1993). Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Osterlind, S. J. (1997). *Constructing test items: Multiple-Choice, Constructed-Response, Performance, and Other Formats*. Boston, MA: Kluwer.

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

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

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

Sireci, S.G., & Ríos, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*, 170-187.

Zwick, R., Donoghue, J., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.