REVISTA DE PSICOLOGÍA Y EDUCACIÓN

Consejo General de la Psicología ESPAÑA

ACIPE

Revista de Psicología y Educación / Journal of Psychology and Education, 2018, 13(1), 23-32 (www.rpye.es)
Doi: http://dx.doi.org/10.23923/rpye2018.01.155 ISSN: 1699-9517

Psychological and Educational Test Score Comparability across Groups in the Presence of Item Bias

Paula Elosua[1] y Ronald K. Hambleton[2]

[1]Univesidad del País Vasco
[2]University of Massachusetts

Abstract: It is common for test publishers to make their most popular educational and psychological tests available in multiple languages and cultures. Occasionally, too, test items are found after publication of these new language versions of tests that may disadvantage members taking these translated tests due to biases. This means that when these tests are used, scores for candidates will be underestimated to some extent and test score validity will be adversely affected. The purpose of this paper is to introduce and demonstrate one possible technical solution to the problem—it involves both differential item functioning (DIF) analysis and statistically equating of test scores. This solution involves two steps: First, any DIF items must be identified using one or more of the standard DIF detection procedures in the language or cultural group of interest. Second, after removing DIF items that may be biasing score interpretations from the actual test scoring, a statistical equating between the original test, and the reduced (shortened) test in the second language/cultural group can be carried out. A demonstration of the methodology is provided in the paper along with a discussion of the advantages and disadvantages of the solution.

Keywords: Test score adjustments, Differential item functioning, Test score equating.

Comparación de puntuaciones en Psicología y Educación cuando se ha detectado sesgo

Resumen: En la práctica psicológica y educativa es frecuente administrar los mismos instrumentos de evaluación a diferentes grupos lingüísticos y culturales. Si estudios posteriores a la publicación de un test detectan la presencia de sesgo, pudiera ocurrir que varios ítems perjudicaran a alguno de los grupos. Como consecuencia, las puntuaciones se subestimarían, y la validez quedaría comprometida. El objetivo de este trabajo es plantear una posible solución técnica a este problema. La propuesta se apoya en el análisis del funcionamiento diferencial del ítem (FDI), y en la equiparación estadística de las puntuaciones. El procedimiento se implementa en dos etapas; en primer lugar y utilizando las técnicas de detección de FDI se identifican los ítems sesgados y en segundo lugar, tras eliminar su efecto sobre la interpretación de las puntuaciones, se procede a equiparar estadísticamente la prueba original y su versión reducida (acortada). En este artículo se muestra cómo llevar a cabo este ajuste y se discuten las ventajas e inconvenientes de esta solución.

Palabras clave: Ajuste de puntuaciones, Funcionamiento diferencial del ítem, Equiparación de puntuaciones.

It is common for test publishers to make their most popular educational and psychological tests available in multiple languages and cultures. It has been reported, for example, that a number of popular American intelligence and personality tests are now available for use in more than 50 languages and cultures (Elosua & Iliescu, 2012). Occasionally, too, after publication of these new language versions of tests, items are found that may disadvantage members taking these translated tests due to biases—for example, a concept may be unknown or a strange word introduced in the translation process, and it went undetected in the review process. This can happen because these translations are often validated with

Received: 10/10/2017 - Accepted: 29/11/2017 - Avaliable online: 14/12/2017

*Correspondence: Paula Elosua.
Universidad del País Vasco
Avda. Tolosa, 70.
Postcode: 20018, San Sebastian, España.
E-mail: paula.elosua@ehu.es

Elosua, P., & Hambleton, R. K. (2018). Psychological and Educational Test Score Comparability across Groups in the Presence of Item Bias. *Journal of Psychology and Education, 13*(1), 23-32, doi: http://dx.doi.org/10.23923/rpye2018.01.155

only judgmental reviews. According to the ITC Guidelines for Test Adaptation (Hambleton, Merenda, & Spielberger, 2005; Muñiz, Elosua, & Hambleton, 2013), empirical analyses should be carried out too but often this step is skipped in the validation process due to financial and/or time constraints. This means that when the tests are used, scores for candidates will be underestimated to some extent and test score validity will be adversely affected.

The purpose of this paper is to demonstrate one possible technical solution to the problem—it involves both differential item functioning (DIF) analysis and statistically equating of test scores. This solution contains two steps: First, any DIF items must be identified using one or more of the standard DIF detection procedures in the language/cultural group of interest (for a review of procedures, see Penfield & Camilli, 2007). Second, after removing DIF items that may be biasing score interpretations, a statistical equating between the original test, and the reduced test in the second language/cultural group can be carried out (Kolen & Brennan, 2004). A demonstration of the methodology is provided in the paper along with a discussion of the advantages and disadvantages of the solution.

The issue of DIF continues to be an important topic in the field of psychological and educational testing. DIF means that an item doesn't perform in the same way for different subgroups of people that have the same level or score on the measured construct; there is an interaction between the characteristics of the item and the subgroup characteristics, and this interaction has a significant effect on the item psychometrics properties. The item doesn't satisfy the invariance property in two samples and the lack of this property being satisfied in the data is one threat to the validity of the test scores (see, Millsap, 2011). If one item presents DIF the expected score for this item is different for people that have the same score on the construct but they are belonging to different groups (e.g., racial, ethnic, gender).

The detection of DIF may be carried out in the early phase of test construction before any field testing by using judges' reviews (Elosua, Mujika, Almeida, & Hermosilla, 2014). This is a good feature in any test development project. But it is not always done. This makes the empirical work carried out on field-test data or possibly operational test data even more important.

Statistical work involves statistical tools under the null hypothesis of equivalence of the parameters of the item among groups. A lot of procedures have been developed for this purpose: Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), item response theory (Hambleton, Swaminathan, & Rogers, 1991), the standardized mean difference (Zwick & Thayer, 1996) and many more. After statistical detection a content analysis would be desirable in order to determine the causes of the differential functioning (Elosua & López, 2007; Zenisky, Hambleton, & Robin, 2003).

Most of the studies of DIF are performed in the item level, detecting aberrant items in the test construction phase or analyzing the behaviour of items and the responses they elicit. If the DIF detection is carried out during the process of test construction, the test developers will usually delete the flagged items from the test when there are good reasons for doing so. In this framework the interest is not the impact of the DIF items in the total sore performance, but the ultimate the construction of unbiased tests. However in many situations, for example, with a translated version of a test from English to Spanish, it may be impossible to revise the Spanish version of the test. It has been printed, and is being used, and it is only after a lot of data have been collected that the DIF can even be detected. In this situation, the focus is not so much on the DIF at the item level, but the practical consequence of item level bias, perhaps in several items, on the uses of scores in the translated tests. Unlike in the test construction phase, in this context, it is not possible to remove items, at least until the next edition of the test is prepared and this could take many years.

What to do then? Any test score interpretations with the target language version of the test should be completed with information about the impact of the DIF items on the total score. In this framework, the information regarding the number of DIF items is not especially helpful but the impact of those DIF items along the test score continuum certainly is, and needs to be estimated. Under the hypothesis that the two test forms are structurally equivalence (i.e., have the same factor structure) and the differences between them are focused on the measurement equivalence or scalar equivalence (van de Vijver & Poortinga, 1997), it would be convenient to find a statistical way to adjust scores for those differences.

Suppose we are interested in the validity of numerical aptitude tests among English and Spanish groups. The detection of DIF at the item level would give us important information regarding the relationship of each item and the measurement construct. This information could be used to analyze the cognitive aspects of the items, and would give us information about the characteristics of the item that have differential impact on the language group of the students. But, the psychologist who is using these tests could be more interested in the information regarding the differential test functioning (DTF), and its effect on the scores. The interest would be centred in getting the equivalence scores between groups under the assumption that the two forms of the same test have been used. This information could be used to obtain comparable scores between groups.

This approach analyzes jointly two aspects of measurement invariance: differential item functioning and test equating. The purpose of equating would be to allow the comparability of scores obtained by means of different forms of a test and in different circumstances (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004). In the context of DIF the aim would to get the comparability between forms of the same tests in different groups assuming the DIF is present. Using equating techniques after DIF detection could adjust the scores obtained on tests that measure the same construct.

In this context the goal of this work was to show one complete process to get the comparability of the scores between groups using one test that presents DIF items in the framework of item response theory, and also to show the validity of test characteristic function concept in assessing differential test functioning. Other approaches are possible too (Elosua & López-Jauregui, 2008), but the IRT offers us one tool to derive and understanding the process going from the differential functioning in the item level, to the differential functioning at the test level. The research was carried out in two stages. First, a study of differential item functioning was performed based on the application of non parametric procedures for detecting DIF. The detection of DIF is done prior and is necessary before linking metrics in the framework of item response theory. Second, after the estimation and linking of parameters, we estimated the item characteristic functions (ICC) and test characteristic functions (TCF). The firsts one offered graphical and numerical information about DIF at the item level, and the TCF allowed us to get comparable test scores and gave us one graphical as numerical information about the differential test functioning.

The differential test functioning was defined using the differences between the two characteristic functions estimated in two samples. The analysis of the differences on the tests level allowed us to set the possibility of DIF cancellation. If the DIF cancellation is observed (e.g., one item might favor one group, and a second DIF item might favor the other), there might not need to be any score adjustment. But, if conditional score differences along the test score continuum continue to exist, then adjustments would be needed.

METHOD

DIF DETECTION PROCEDURES

→ *Standardized Mean Difference* (SMD) (Zwick & Thayer, 1996) is an extension of the formulation of Dorans and Holland (1993) who proposed a DIF indicator that is the conditional difference between the means of the reference and focal groups. This statistician calculates the difference between the grand mean of the items scores for the focal group and the mean item score for the reference group, "standardized" as if the reference group distribution across levels were the same as the focal group distribution. The value of this statistic depends on the scale of response, and in order to obtain one index independent on the scale, the authors proposed to divide it by the standard deviation of the focal group and reference group combined.

→ *Mantel*. This procedure compares the mean obtained in one item for persons belonging to two groups that have the same score level (Spray & Miller, 1994; Zwick, Donogue, & Grima, 1993). The Mantel statistic estimates the interaction between group/item and it follows the chi-square distribution with one degree of freedom.

The size of the effect can be analyzed with the Standardized Mean Difference. (SMD; Dorans, & Kulick, 1986) divided by the standard deviation of the combination of the referemce and focal groups. Following the criterion used by the ETS, an item presents moderate DIF when besides the statistical significance of the used statistician, the size of the effect is greater or equal to 0.17 and minor or equal to 0.25. One item presents severe DIF if the effect size is greater than 0.25.

→ *Differences between Item Characteristic Functions (ICF)*. This method compares the expected scores functions obtained in two different groups after the metrics have been equated. In absence of DIF and because the invariance property of item response theory models, is assumed that the estimated Item Characteristics Functions will be equivalent. The difference between ICF will be indicator of presence of DIF. It is possible to derive one measure of the difference by taking account of the differences in several values among the theta continuous or taking account the area between two ICFs (Kim & Cohen, 1991; Raju, 1988; Raju, van der Linden, & Fleer, 1995). In this work we estimated the differences in the expected score across 10 points on the ability metric by adding the square of the differences in each point.

→ *Differential Test Functioning*. The differential test functioning is assessed by the analysis of the differences between the test characteristic function estimated in two samples over all items.

PARTICIPANTS

The sample for the study was comprised of 1328 participants belonging to two language groups, named group A (called the "source group") and group B (called the "target group"). The first one corresponded to the population to which the original test was developed (*N* = 967), and the second was the group in the language for which the test was adapted (*N* = 361).

INSTRUMENT

The data came from a popular personality test. The scale consists of 65 items with 5 ordered category response options (scored 1 to 5). The score range was from 65 to 325.

RESULTS

DESCRIPTIVE STATISTICS

The sample statistics and coefficient alpha for each sample are reported in Table 1. The equality of the variance test was significant ($F_{966,360}$ = 1.31; *p* = .002). The difference between the means was statistically significant ($t_{734.275}$ = 10.66; *p* < .001) and the effect size was .07. Coefficient alpha obtained in sample A was .922, and the value in sample B was .910. The equivalence between the two internal consistency coefficients was assessed with the statistic proposed by Feldt (1969). The value obtained allowed us to accept the hypothesis of equivalence between coefficients (*w* = .86; *p* = .960).

Table 1
Descriptive Statistics for the Two Groups in the Personality Test

Group	N	Minimum	Maximum	Mean	SD	Cronbach's alpha
A	967	125	289	205.40	26.7	.922
B	361	161	286	221.34	23.3	.910

Table 2
Differential Item Functioning Analysis Results

Item	Mantel	SMD/S_i	DIF.E$[X]_i$	Item	Mantel	SMD/S_i	DIF.E$[X]_i$
1	64.438	-0.552	0.358	33	7.977	0.016	0.081
2	8.878	0.172	0.783	34	0.295	0.095	0.255
3	40.828	-0.404	0.986	35	3.615	0.166	0.498
4	0.407	-0.040	0.15	36	6.582	0.663	4.905
5	5.488	-0.152	0.17	37	115.742	-0.103	0.529
6	51.277	0.391	1.981	38	3.413	-0.199	1.26
7	0.129	-0.025	0.091	39	15.453	-0.635	3.357
8	25.861	0.271	1.551	40	131.714	0.077	0.441
9	15.642	-0.227	0.46	41	2.257	0.458	1.971
10	2.864	0.105	0.288	42	65.484	-0.147	0.524
11	80.850	0.552	4.676	43	9.390	0.022	0.117
12	10.046	-0.187	4.649	44	0.174	-0.162	0.057
13	63.593	0.458	3.11	45	6.306	0.100	0.355
14	9.642	-0.184	0.069	46	3.849	0.307	1.078
15	6.110	-0.157	0.462	47	23.280	-0.054	0.033
16	10.609	-0.167	0.058	48	0.958	0.022	0.642
17	17.152	0.230	1.367	49	0.213	-0.138	0.113
18	12.154	0.198	1.001	50	5.047	-0.181	0.077
19	6.535	0.133	0.445	51	7.555	-0.110	0.355
20	2.491	0.095	0.236	52	3.948	-0.283	0.333
21	2.568	-0.092	0.117	53	27.420	0.197	0.532
22	23.822	0.287	1.267	54	11.229	0.055	0.102
23	0.092	-0.010	0.156	55	1.116	-0.034	0.092
24	155.654	-0.755	4.309	56	0.429	0.230	1.602
25	6.480	-0.193	0.501	57	13.236	0.133	0.541
26	9.842	0.157	0.974	58	6.730	-0.097	0.2
27	0.511	0.043	0.31	59	4.202	-0.231	0.536
28	17.284	-0.237	0.12	60	12.067	0.232	0.996
29	14.777	0.222	0.774	61	19.630	0.088	0.649
30	17.407	0.234	1.218	62	1.727	-0.639	3.659
31	3.647	0.086	1.071	63	108.265	0.029	0.205
32	8.846	-0.139	0.409	64	0.138	0.171	1.134
				65	13.704	0.016	0.081

DIF ANALYSIS

The results of the Mantel and the Standardized Mean Difference are showed in Table 2. According to the results and the joint criteria of significance (p<.001) and effect size, 30 items showed differential item functioning; that is 46.15% of the total items. 16 of them favored to the B sample or focal sample, and the rest, 14 items, showed differential item functioning against this group.

ESTIMATION OF THE IRT MODEL

The graded response model was estimated (Samejima, 1997) using PARSCALE 4.1 by fixing the distributions of theta scores to have a mean of 0.0 and a standard deviation of 1.0. The estimation was carried out independently in the two samples and after the metrics were equated using a linear equating procedure. The selection of the anchor items was made using two criteria: the absence of DIF and the difference between estimated thresholds. The first criterion allowed to define one set of anchor items showing no DIF (i.e., invariance of groups), and the second criterion allowed to control the estimation error of the parameters related with the size of samples. The maximum difference between estimated thresholds was fixed at 1.0. The estimation of the linear equating parameters, slope and intercept, were 1.06 and 0.33 respectively. After equating, the mean and standard deviation of the theta estimated in the B sample were 0.33 and 1.06. Group B should higher scores on the test, and was about as variable in score distribution as Group A.

ITEM CHARACTERISTIC FUNCTIONS

The item characteristic functions were estimated for each item in each sample. In the presence of DIF there will be differences between two functions. This procedure gave graphical information about the presence/absence of DIF in the all range of theta. Figure 1 shows the ICF for two items with DIF (items 37 and 24), and another two without DIF (items 4 and 48). This graphical information can be extended with the numerical index. For all the items of the test the estimated ICF difference values can be read in Table 2. The correlation between this index and the Mantel statistics was 0.80.

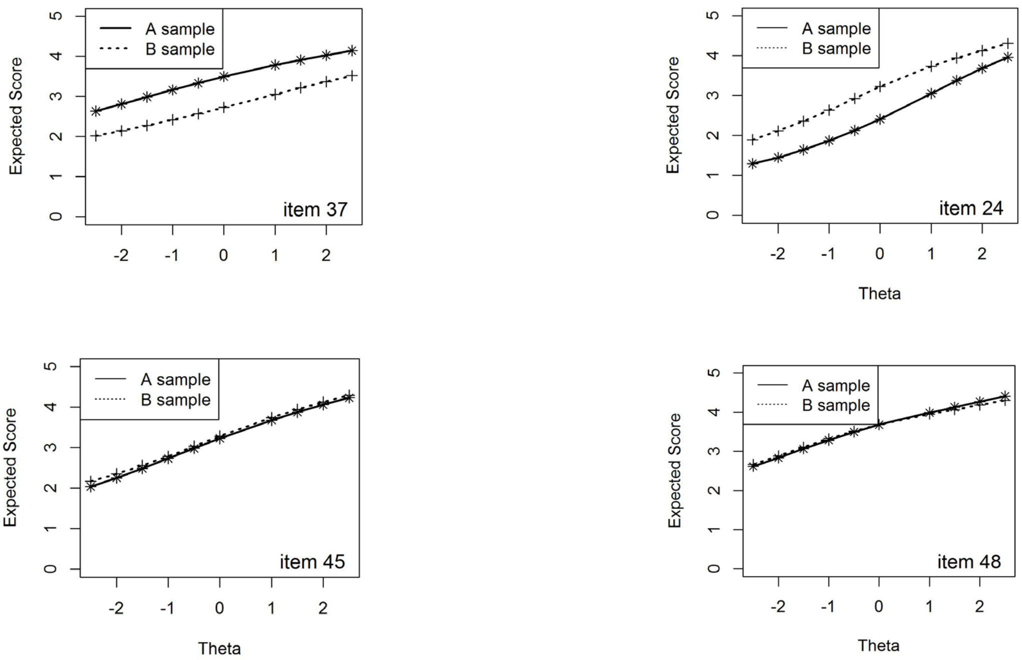

Figure 1. Item characteristics functions.

TEST CHARACTERISTIC FUNCTION

The sum of the expected scores among items in the range of theta gave the value of the expected score on the total tests. If the tests were parallel, the two tests characteristic functions might be overlapped and the index of differential test functioning would be 0. But the graphical representation didn't show two overlapped functions (see Figure 2) , and neither was the value of the differential test functioning at 0.0. The value of 61.75 was obtained for this differential test functioning index.

The Figure 3 shows the differences between the expected total score across the range of theta. The value of the difference was close to 0 at the -2.0 value of theta, but this difference increased and remained constant across the level of theta greater than -0.5, to a 5 point reestimated using only items without DIF the differences in the expected scores became very small. Figure 4 shows those new test characteristic functions, and Figure 3 represents the graph of the differences over theta. The index for the differential test functioning has been reduced; the new index was 13.80. The mean value of the differences was now only 1.3 points. Those new Test Characteristic Functions include the information for score comparability when comparability is the objective of the assessment.

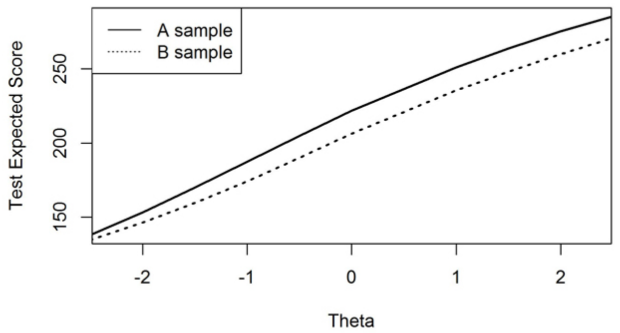

Figure 2. Test characteristic functions for the two forms.

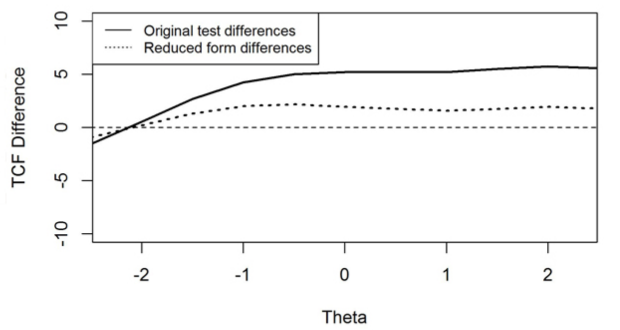

Figure 3. Differences among the expected scores over the ability scale.

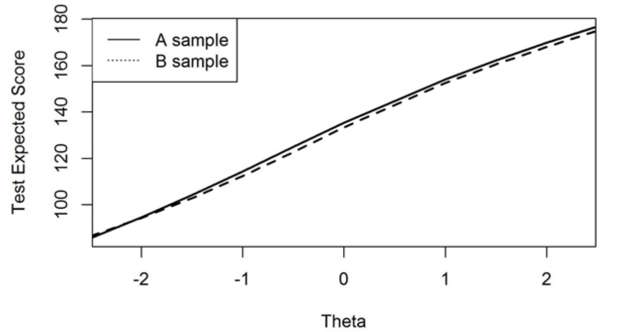

Figure 4. Test characteristic functions without DIF.

DISCUSSION

DIF analyses are routine in test development projects, and they have the desirable effect of enhancing test score validity, because problematic items can be removed. The problem is more consequential when the test itself cannot be changed, at least until the next edition of a test is produced. With many tests, this time period could extend beyond five years—a long time to be using tests with biases. In practice, and when the problems are identified, probably users make their own subjective judgments. But this level of subjectivity results in a loss of standardization in the testing process. These problems might arise when a test is developed but DIF analyses are not carried out. It is only later the problem is identified and then the bias is present. The much more common situation occurs when tests from one language and culture are adapted for use in other languages and cultures. The tests are published prior to DIF studies, and undetected DIF can be a major threat to test score validity.

The procedure we recommend starts with the detection of DIF based on the total score or other procedures in order to get information about the DIF items. After the IRT model is estimated (we used the graded response model), the score scales are equated using linear equating (or another suitable equating method) using all the items as an anchor except those showing some level of DIF. The process continues through the estimation of the ICFs in the two samples. The method of ICF gives to the research very important graphical information about the amount, sign and distribution over theta of the DIF, and also, it is possible to derive one numerical index based in the differences founded between two ICFs. Starting with the item characteristic function, it is simple to develop one measure of differential functioning at the test level. All the information that the research needs in order to determine the amount of differential test functioning, and also to linking scores between samples is contained in the two test characteristic functions. We showed the differences in those numerical indexes and graphical outputs between one test with DIF, and the same test after removing DIF items.

The approach described in this paper treats jointly two aspects regarding to the invariance of the measurement in the context of psychological and educational testing; the differential item functioning and the equating of scores. In this paper two equating processes were carried out; the first was made in the estimation of the model in order to place the metric of the parameters on the same scale. The second one is related to the final results of the process, and it was intended to get the comparability of the scores.

The finding from this study is that a two-stage approach to the problem is practical, and can have a consequential impact on the interpretation of scores. We would be supportive of our approach being used whenever tests that have important consequences for examinees have not been studied for DIF. DIF studies are not perfect—items can still be missed in an analysis (a type II error), and also the IRT portion of the approach requires experienced persons to do the actual equating, and sufficient data to carry out the equating analyses.

In thinking about subsequent research, we would begin with more DIF analyses on other translated tests, or tests developed within a language group where DIF studies were not carried out. It seems important to identify the size of the problem in educational and psychologically tests. After all, it is not that common to find DIF items, and even when it is found, often it tends to balance out—some items show DIF against males and others against females, or against the source language or the target language group, and the overall impact can be near zero. DIF due to cultural and language issues, or poor translations, have been less studied and so it is here we might try to learn more about the impact of DIF.

It is clear that for at least some tests, this two stage process is going to be relevant. Then the next question would surely concern best DIF methods and perhaps statistical significance tests for the best ones. Oshima, Raju and Nanda (2006) have looked at this problem and others too in the medical testing literature. But it is a worthy problem for study. Another good topic concerns the impact of item deletion on a test and its impact on reliability and validity of scores. No one would advocate using a biased test, but then if too many items are deleted, questions arise about the reliability and validity of scores based on a reduced set of items, and even the comparability of the source and language versions of the test. Deleting 3 to 5 items from a 50 achievement test seems worth doing to eliminate the bias, but eliminate 15 items would raise serious questions about the reliability and validity of the test with a reduced set of items, and it might be much harder to argue that the source and target language versions of a test are structurally equivalent.

In summary, recent years have seen a substantial increase in the number of studies involving measurement invariance or differential item functioning. DIF detection is important in the studying the conditions for good assessment, but from a practical point of view a further step is necessary. It is time to consider the impact of differential item functioning on test level and to offer practitioners correct ways of using tests and interpreting scores in the presence of bias. The solution shown in this paper is based on the application of item response theory models. IRT provides both invariant item statistics and ability estimates but these features will be obtained when there is a reasonable fit between the chosen model and the dataset. Certainly some issues and technical problems remain to be solved in the IRT field but it would seem that item response model technology is more than adequate at this time to serve a variety of uses (Hambleton, 1990).

- Conflict of interest

The authors declare no conflict of interest.

Acknowledgements
This work was partially supported by the University of the Basque Country [GUI15/24] and the Spanish Ministry of Economy and Competitiveness [PSI2014-54010, PSI2017-82539].

REFERENCES

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368. doi: 10.1111/j.1745-3984.1986.tb00255.x

Elosua, P., & Iliescu, D. (2012). Tests in Europe. Where we are and where we should to go. *International Journal of Testing, 12*, 157-175. doi: 10.1080/15305058.2012.657316

Elosua, P., & López, A. (2007). Potential DIF sources in the adaptation of tests. *International Journal of Testing, 7,* 39-52. doi: 10.1080/15305050709336857

Elosua, P., & López-Jáuregui, A. (2008). Equating between linguistically different tests. *Journal of Experimental Education, 76,* 387-402. doi: 10.3200/JEXE.76.4.387-402

Elosua, P., J. Mujika, Almeida, L., & Hermsosilla, D. (2014). Procedimientos analítico-racionales en la adaptación de tests. Adaptación al español de la Batería de Pruebas de Razonamiento. *Revista Latinoamericana de Psicología, 46*(2), 117-126. doi: 10.1016/S0120-0534(14)70015-9

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34,* 363-373.

Hambleton, R. K. (1990). Item Response Theory: Introduction and Bibliography. *Psicothema, 2*, 97-107.

Hambleton, R. K., Merenda, P., & Spielberger, C. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

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

Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*(3), 269-278. doi:10.1177/014662169101500307

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking.* New York: Springer-Verlag.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición. *Psicothema, 25*, 149-155. doi: 10.7334/psicothema2013.24

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*,1-17. doi: 10.1111/j.1745-3984.2006.00001.x

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol 26:* Psychometrics (pp. 125-167). Amsterdam, Netherlands: Elsevier.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502. doi: 10.1007/BF02294403

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368. doi: 10.1177/014662169501900405

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag

Spray, J. A., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously-scored test items*. Iowa City, IA: American College Testing Program.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. doi: 10.1177/0146621616668015

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29-37. doi: 10.1027/1015-5759.13.1.29

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating*. New York: Springer.

Zenisky, A.L., Hambleton, R. K., & Robin, F. (2003). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment,* 9(1&2), 61-78. doi: 10.1080/10627197.2004.9652959

Zwick, R., Donogue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.doi: 10.1111/j.1745-3984.1993.tb00425.x

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187-201. doi: 10.3102/10769986021003187