International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 774–786
https://doi.org/10.21449/ijate.1499099
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Examination of differential item functioning in PISA through univariate and multivariate matching differential item functioning

Ahmet Yıldırım[1*], **Nizamettin Koç**[2]

[1]Ankara Hacı Bayram Veli University, Faculty of Literature, Department of Psychology, Ankara, Türkiye
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Retired, Ankara, Türkiye

ARTICLE HISTORY

Received: June 10, 2024

Accepted: Sep. 02, 2024

Keywords:

Differential item functioning,

Multivariate matching,

Purified matching variable.

Abstract: The present research aims to examine whether the questions in the Program for the International Student Assessment (PISA) 2009 reading literacy instrument display differential item functioning (DIF) among the Turkish, French, and American samples based on univariate and multivariate matching techniques before and after the total score, which is the matching variable, is purified of the items flagged with DIF. The study is a correlational survey model research, and the participants of the study consist of 4459 Turkish, French, and American students who took booklets 1, 3, 4, and 6 in the PISA 2009 reading literacy measure. Univariate and multivariate (bivariate, trivariate, and quadrivariate) DIF analyses were performed through logistic regression before and after purifying the matching variable off the items displaying DIF. Literature was used to detect extra matching variables, and multiple linear regression analysis was carried out. As a result of the analyses, it was discovered that using extra matching variables apart from the total score reduces type I errors. It was also concluded that the exclusion of DIF items (removal of items with DIF) while calculating the total score led to variation in the number of questions detected as DIF and DIF levels of the items, although it did not yield consistent results.

1. INTRODUCTION

Adapting measures developed in linguistic community for use in different communities is a practice frequently used in recent years (Allalouf, Hambleton & Sireci, 1999). The translation of the Binet-Simon Intelligence Test from the original language to the source language can be considered one of the oldest samples of this (Hambleton, 1993; Hambleton & Patsula, 1999). Cross-cultural studies require adaptation of measures and administration in various communities (Van de Vijver & Tanzer, 2004). However, ensuring that the measured structure is equivalent across all cultures is crucial for making meaningful interpretations (Braun & Harkness, 2005; Gierl, 2000).

Recently, there has been a noticeable increase in intercultural evaluation studies conducted internationally, as well as in the number of countries participating in these studies. For example, a total of 65 countries and non-members of the Organization for Economic Co-operation and

*CONTACT: Ahmet YILDIRIM ahmet-yildirim@hbv.edu.tr Ankara Hacı Bayram Veli University, Faculty of Literature, Department of Psychology, Ankara, Türkiye

e-ISSN: 2148-7456

Development (OECD) participated in PISA (Program for International Student Assessment PISA) in 2012, in which Turkey also participated. Similarly, 63 countries got involved in Trends in International Mathematics and Science Study (TIMSS) in 2011 (International Association for the Evaluation of Educational Achievement, 2012). Considering that the countries participating in these studies and the people living in these countries differ in terms of ethnicity, language, and many other variables (Sireci and Rios, 2013) the necessity of adapting the tests developed within the scope of international studies to the language and culture of the participating countries emerges.

In adaptation studies, it is an important validity issue that the instruments adapted are not comparable with the original tests (Arffman, 2010; Ercikan et al., 2004; Perrone, 2006; Sireci & Allalouf, 2003). Because when the scores obtained from the tests are not comparable, it becomes difficult to make comparable interpretations based on the scores of the individuals taking the test in the cross-cultural studies (American Educational Research Association, 2014). PISA is one of the crosscultural studies administered in many different countries. Wealthier countries tend to participate in PISA as they have an assessment culture and also would like to see the trends in their educational system based on time. However, economically disadvantaged countries also started to show interest in large-scale international research so that they can see improvement in their education system. Currently, lower-middle-income countries such as Georgia and Indonesia; and upper-middle-income countries like Bulgaria and Brazil have participated in PISA administrations. As a result, PISA has a huge coverage in terms of participation (Organisation for Economic Co-operation and Development, 2015). The aim of PISA is to determine the competencies of 15-year-old students in three main areas: (a) reading skills, (b) mathematics, and (c) science literacy. Regardless of the construct measured by the test, there are basically two factors that affect the equivalence of measurement instruments used in international studies such as PISA: (1) translation, (2) culture (Gradshtein, Mead & Gibby, 2010).

As the utilization of tests in making important education-related decisions increases and legal issues concerning the use of tests arise, differential item functioning (DIF) and item bias may become an important problem in the evaluation of test validity (Hambleton, Clauser, Mazor & Jones, 1993). Bias causes systematic errors that deform the outcomes acquired from the measures and the evaluation based on these findings (Gierl, Rogers & Klinger, 1999). As testing and testing practices have come to public attention in recent years, test publishers and experts who use tests have to provide evidence that the tests they use and publish are not biased against minorities and are invariant for all participant groups (Hambleton et al., 1991).

Recently, DIF analyses have been frequently utilised to detect items that are not comparable across different communities (Allalouf et al, 1999; Allalouf & Sireci, 1998; Gierl et al., 1999; Gierl & Khaliq, 2000). DIF analyses are used to determine whether the test items function similarly across different groups (Hambleton et al., 1993; Sireci & Swaminathan, 1996; Zumbo, 1999; Zumbo, 2007).

DIF refers to the psychometric difference in how a question functions for two different groups. In other words, DIF can be defined as the distinction in performance between the groups compared concerning the relevant item (Allalouf et al., 1999; Dorans & Holland, 1993). DIF happens when a question in a test works inequivalently for various groups (Clauser & Mazor, 1998; Furlow et al., 2009). The reasons that make it necessary to conduct DIF studies are (Zumbo, 2007): (1) ensuring equity and fairness in assessment and evaluation, (2) Eliminating possible threats for validity, (3) Examining the equivalence of translated tests.

In DIF analyses, individuals in different groups are matched based on a matching variable and contrasted with regard to their performance on items (Camilli, 1992). The determination of a valid and justifiable matching variable is important for obtaining precise results in DIF analyses (Gierl et al., 2000). In DIF analyses, the sum of the item scores (endogenous variable) is usually

employed as the matching variable (Hambleton et al., 1993; Sireci & Rios, 2013). How valid and reliable such matching will be is a question that needs to be answered. It is suggested that matching should be based on an external variable with previously established validity (Gierl, 2004). Unfortunately, such a variable may not always be available (Clauser & Mazor, 1998). The use of additional matching variables should be considered when other variables are thought to be related to the construct or affect individuals' performance on the construct being measured (Sireci & Rios, 2013).

When the secondary factors that lead to the emergence of DIF are elements of the construct assessed by the measure and are consciously measured, these factors are referred to as auxiliary factors. However, when these factors are measured even though they are not components of the construct assessed by the instrument, they are called confounding factors (Boughton et al., 2000; Camilli, 1992; Gierl & Khaliq, 2000). DIF led by auxiliary factors is called benign DIF, while DIF led by confounding factors is called malignant DIF (Boughton et al., 2000; Gierl, 2004). DIF analyses based on multivariate matching provide a better understanding of the causes of DIF and reduce the likelihood of making type I errors (Roussos & Stout, 1996). Within the framework of DIF, the type I error is the detection of an item with DIF when in reality the item does not display DIF (Jodoin, 1999). Determining a reliable and error-free matching variable is critical for obtaining accurate results in DIF studies. Whether the matching variable should be purified of the items with DIF is an important question to be answered in DIF analyses (Sireci & Rios, 2013). The involvement of DIF items in the total score while calculating the matching variable calls into question the appropriateness of the matching variable (Gierl et al., 2000). When conducting DIF analyses, the matching variable needs to be purified. In other words, items labeled as DIF should be discarded and the total score should be recomputed. This recomputed total score is employed as the matching variable for the second logistic regression analysis (Zumbo, 1999). French and Maller (2007) state that the involvement of DIF items in the total score in DIF detection may lead to errors. To control these errors, researchers (French & Maller, 2007; Gierl et al., 2000; Khalid & Glas, 2013; Zumbo, 1999) argue that the total score, which is the main matching variable, should be purified. According to Lee and Geisinger (2016), the purification of the matching variable involves the exclusion of items defined as DIF in the initial DIF analysis when calculating the total score, to put it another way, the use of only non-DIF items when calculating the matching variable (when calculating the total score). Two approaches are adopted in the purification of the matching variable. One of these is the two-stage purification approach and the other is the iterative purification approach. When a single DIF study is conducted to exclude DIF items from the calculation of the matching variable, it is referred to as the two-stage purification approach. If iterative DIF analyses are performed until no items are identified as DIF, it is known as the iterative purification approach (Lee & Geisinger, 2016).

As PISA is an intercultural evaluation study, both English and French versions of all measures used within the scope of PISA are developed, and these tools are sent to the participating countries for adaptation procedures. The two forms of the test are developed in parallel and in this way, it is planned to minimize cultural dependency. As a result of the adaptation, the various language forms of the test are considered to be the same. However, it needs to be demonstrated whether this is the case in reality. Moreover, in DIF studies conducted on items of international tests such as PISA, individuals are usually matched using a single matching variable (total scores) and analyses are conducted in this way. In addition, DIF analyses are conducted without purifying the total score which is the matching variable of the items with DIF. Considering that other variables such as socioeconomic status, parental level of education, home possessions, etc. in addition to individuals' total scores may explain performance differences it is necessary to use other matching variables apart from the total score and to purify the total score of the items with DIF in DIF studies. However, DIF studies are conducted by ignoring the aforementioned properties. They are either conducted by using a single

matching variable such as total score, or they are performed based on the total score including the items tagged with DIF. These might be considered sources of errors in DIF studies. Considering all these problems and drawbacks in DIF studies may lead to erroneous implications, the current study employing purified total score and other matching variables apart from the total score was conducted. As a result, this study was required to examine the effect of using other matching variables such as maternal education level, paternal education level and home possessions in addition to the total score in DIF studies and the effect of purified matching variable on DIF determination.

The general purpose of this study is to determine whether the items in the reading literacy test of PISA 2009 display DIF between the samples of Turkey and the USA by using univariate and multivariate matching methods (before and after purifying the total score of the items with DIF). Within this general purpose, answers to the following research questions were sought:

1. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the univariate logistic regression technique before purifying the total score of the items with DIF?
2. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the multivariate logistic regression technique before purifying the total score of the items with DIF?
3. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the univariate logistic regression technique after purifying the total score of the items with DIF?
4. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the multivariate logistic regression technique after purifying the total score of the items with DIF?

2. METHOD

This study, which aims to identify if the items in the PISA 2009 reading skills instrument display DIF between Turkish and US samples by using univariate and multivariate matching methods is a type of correlational survey research design (Tabachnick & Fidell, 2013). Correlational survey design is used to determine the existence of co-variation between two or more variables (Karasar, 2011).

2.1. Sample

The population of PISA includes students in the age group of 15 in each participating country. In participating countries, the target population includes all students between the ages of 15 years and 3 months and 16 years and 2 months who are attending school. The sampling strategy of PISA is a two-stage stratified sampling. In the first stage, schools with students in the age group of 15 are selected. In the second stage, students are drawn from the sampled schools (Organisation for Economic Co-operation and Development, 2014). Within the framework of this research, studies were performed on the booklets numbered 1, 3, 4, and 6, in which the OECD has revealed the largest number of items, and the Turkish and US samples who responded to the items in these booklets. The Turkish sample includes 1533 students while the US sample includes 1611 students.

2.2. Obtaining Data

The data for this research includes the responses of Turkish and U.S. students to nine items from booklets 1, 3, 4, and 6 of the PISA 2009 reading literacy test, which contained the highest number of items released by the OECD. The data were accessed from the official page of the OECD (http://www.oecd.org/pisa/data/). Six of the nine items in the booklets were selected-response and three were constructed-response. Constructed-response items are dichotomous items that are scored 1-0. For that reason, open-ended items do not have partial scores.

2.3. Data Analysis

2.3.1. *Testing dimensionality*

It is argued that the multidimensionality of items leads to DIF. For this reason, unidimensionality is a requirement for DIF identification methods that require unidimensionality (Wen, 2014). Confirmatory factor analysis was utilized to test dimensionality and the results are shown in Table 1.

Table 1. *Goodness of fit measures estimated from Turkish and US samples.*

Indices of goodness of fit	Turkish Sample	US Sample
χ^2/df	1.328	1.948
CFI	.991	.987
GFI	.995	.992
RMSEA	.015	.024

The results estimated based on confirmatory factor analysis support the unidimensionality assumption. In other words, the unidimensional factor model fits the reading literacy data of Turkey excellently, and the USA as seen in Table 1 (Hu & Bentler, 1999; Mcdonald & Ringo Ho, 2002). It could be stated that the factor structure of the reading literacy test is invariant across language groups.

2.3.2. *DIF detection technique*

In this study, logistic regression was used as a DIF detection technique. In logistic regression analysis used to determine DIF, variables are included in the model hierarchically. "In Step 1, the matching variable is introduced into the model as an independent variable. In Step 2, the group variable is added. In Step 3, the interaction term is incorporated into the equation. In logistic regression, the chi-square test is used to assess statistical significance, and the contribution of each variable to the model is evaluated. The chi-square value from the first model is then subtracted from the value obtained in the third model. The chi-square value obtained is compared with the chi-square distribution with 2 degrees of freedom. Degrees of freedom 2 is calculated by subtracting the degrees of freedom in the first model (1) from the degrees of freedom in the third model (3) (Crane et al, 2006; Gierl et al, 2000; Hidalgo & Lopez-Pina, 2004; Jodoin, 1999; Sireci & Rios, 2013; Zheng et al., 2007). The result obtained by subtracting the R^2 value obtained from the third model from the R^2 value obtained from the first model provides evidence for the effect size of DIF (Sireci and Rios, 2013; Zumbo, 1999). Logistic regression can also be applied when more than one variable is used to match individuals (Sireci & Rios, 2013). Nagelkerke R^2 value can be employed as an effect size to determine the magnitude of DIF. In order to claim that there is a DIF, the difference in R^2 values between models should be at least .13 (Zumbo, 1999). Zumbo and Thomas (1997) suggested the cut-off points in Table 2 for $\Delta R^2 = R^2$ (M3) - R^2 (M1) to be used in interpreting the magnitude of DIF for logistic regression (cited in Hidalgo and Lopez-Pina, 2004).

Table 2. *Cutt-of points for logistic regression ΔR^2 value.*

ΔR^2	*DIF level*
$\Delta R^2 < 0.13$	A level DIF (No DIF or might be neglected).
$0.13 \leq \Delta R^2 < 0.26$	B level DIF (Moderate DIF).
$\Delta R^2 \geq 0.26$	C level DIF (Serious DIF).

2.3.3. *Detection of additional matching variables*

A literature review was conducted to determine matching variables that may be related to reading skills in addition to the total score. Later on, multiple linear regression was carried out to determine the variables of which regression coefficients are significant. The results belonging to multiple linear regression are presented in Table 3.

Table 3. *Variables and regression coefficients based on multiple linear regression analysis.*

Variables	Regression coefficients	
	β	Standardised Beta
Maternal education level	.21	.17*
Paternal education level	.17	.13*
Attitude towards school	.04	.02
Home possessions	.21	.10*
Family wealth	.03	.01

*$p<0.05$

Table 3 indicates that maternal education level, paternal education level, and home possessions are significant indicators of reading literacy. For this reason, these three variables were considered additional matching variables, alongside the total score on the reading literacy test.

3. RESULTS

This section presents the findings obtained in line with the sub-questions of the study. The findings obtained from univariate and multivariate matching-based DIF analyses conducted before and after the purifying the total score of DIF items were compared.

3.1. Results Regarding Univariate DIF Before Purification

Table 4 indicates the logistic regression-based univariate DIF analysis performed before purifying the total score. Table 4 indicates that four of the nine items display significant DIF between the Turkish and US samples. The results reveal that all 4 items contain DIF at level A.

Table 4. *DIF results based on univariate matching.*

Item Number	(ΔR^2)	DIF Level
R414Q02	.008*	A
R414Q06	.004*	A
R414Q09	.003	
R414Q11	.006*	A
R452Q03	.003	
R452Q04	.001	
R452Q07	.004*	A
R458Q01	.003	
R458Q07	.000	

*$p<0.05$

3.2. Results Regarding Multivariate DIF Before Purification

3.2.1. *Bivariate DIF analysis*

Table 5 indicates the logistic regression-based bivariate DIF analysis performed before purifying the total score. Based on Table 5, four of the nine items displayed significant DIF between the Turkey sample and the US sample. The results reveal that all four items contain DIF at level A. In addition, when compared to univariate DIF analysis, the use of the maternal education level variable apart from the total score did not lead to any change in the number of items labeled as having DIF.

Table 5. *DIF results based on bivariate matching (total score plus maternal education level).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.004*	A
R414Q06	.004*	A
R414Q09	.001	
R414Q11	.006*	A
R452Q03	.003	
R452Q04	.002	
R452Q07	.005*	A
R458Q01	.002	
R458Q07	.001	

*$p<0.05$

3.2.2. *Trivariate DIF analysis*

Table 6 indicates the logistic regression-based trivariate DIF analysis performed before purifying the total score.

Table 6. *DIF results based on trivariate matching (total score plus maternal education level plus paternal education level).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.004	
R414Q06	.004	
R414Q09	.001	
R414Q11	.006*	A
R452Q03	.003	
R452Q04	.003	
R452Q07	.006*	A
R458Q01	.003	
R458Q07	.002	

*$p<0.05$

Table 6 indicates that two of the nine items show a significant DIF between the Turkish and US samples. The results reveal that both items show level A DIF. Compared to the univariate DIF analyses, the use of the variables of maternal education level and paternal education level in addition to the total score lessened the number of items labeled as DIF from four to two.

3.2.3. *Quadrivariate DIF analysis*

Table 7 shows the logistic regression-based quadrivariate DIF analysis performed before purifying the total score.

Table 7. *DIF results based on quadrivariate matching (total score plus maternal education level plus paternal education level plus home possessions).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.004	
R414Q06	.002	
R414Q09	.002	
R414Q11	.006	
R452Q03	.003	
R452Q04	.004	
R452Q07	.006	
R458Q01	.003	
R458Q07	.003	

According to Table 7, no item displayed DIF between the Turkish and US samples. As a result, compared to univariate DIF analyses, the use of other predictor variables apart from the total score reduced the number of items labeled as DIF from four to zero.

3.3. Results Regarding Univariate DIF After Purification

Table 8 indicates the logistic regression-based univariate DIF analysis performed after purifying the total score. According to Table 8, three of the nine items displayed significant DIF between the Turkish sample and the US sample. The results show that all three items contain DIF at level A. It is seen that purifying the total score off the items with DIF reduced the number of items flagged with DIF into three.

Table 8. *DIF results based on univariate matching.*

Item Number	(ΔR^2)	DIF Level
R414Q02	.018*	A
R414Q06	.003	
R414Q09	.002	
R414Q11	.005*	A
R452Q03	.002	
R452Q04	.001	
R452Q07	.003*	A
R458Q01	.002	
R458Q07	.001	

*$p<0.05$

3.4. Results Regarding Multivariate DIF After Purification

3.4.1. *Bivariate DIF analysis*

Table 9 indicates the logistic regression-based bivariate DIF analysis performed after purifying the total score. According to Table 9, three of the nine items displayed significant DIF between the Turkey sample and the US sample. The results demonstrate that all three items contain DIF at level A. Moreover, when compared with the univariate DIF analysis, the number of the items tagged with DIF remained the same.

Table 9. *DIF results based on bivariate matching (purified total score plus maternal education level).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.005*	A
R414Q06	.008*	A
R414Q09	.001	
R414Q11	.005*	A
R452Q03	.003	
R452Q04	.001	
R452Q07	.004	
R458Q01	.001	
R458Q07	.002	

*$p<0.05$

3.4.2. *Trivariate DIF analysis*

Table 10 indicates the logistic regression-based trivariate DIF analysis performed after purifying the total score. According to Table 10, two of the nine items displayed significant DIF between the Turkish sample and the US. The results reveal that both items contain DIF at level A. Compared to the univariate DIF analysis, the use of maternal education level and

paternal education level variables in addition to the adjusted total score decreased the number of items labelled as DIF from three to two.

Table 10. *DIF results based on trivariate matching (purified total score plus maternal education level plus paternal education level).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.005	
R414Q06	.007*	A
R414Q09	.002	
R414Q11	.005*	A
R452Q03	.003	
R452Q04	.002	
R452Q07	.004	
R458Q01	.003	
R458Q07	.002	

*$p<0.05$

3.4.3. *Quadrivariate DIF analysis*

Table 11 demonstrates the logistic regression-based quadrivariate DIF analysis performed after purifying the total score.

Table 11. *DIF results based on quadrivariate matching (purified total score plus maternal education level plus paternal education level plus home possessions).*

Item Number	(ΔR^2)	DIF Level
R414Q02	.006	
R414Q06	.004	
R414Q09	.003	
R414Q11	.004	
R452Q03	.003	
R452Q04	.004	
R452Q07	.004	
R458Q01	.003	
R458Q07	.002	

Table 11 shows that none of the nine items were tagged with DIF between the Turkish sample and the US sample. When compared with univariate DIF analyses, it is seen that the use of other predictor variables apart from the purified total score reduced the number of items labeled as DIF from three to zero.

4. DISCUSSION and CONCLUSION

It was found that the use of other matching variables apart from the total score led to a decrease in the number of DIF items in general. When the univariate matching method was used, while four items were labeled as having DIF between Turkish and US students using the univariate matching method, none of the items were labeled as having DIF in the DIF analysis based on four-variable matching. Based on this point, it can be argued that additional matching variables explain the DIF displayed by the items in univariate DIF analyses and lead to a reduction in the first type error. This finding is compatible with the findings of studies (Arıkan et al., 2018; Çet, 2006; Roussos & Stout, 1996; Yıldırım & Yıldırım, 2011; Yılmaz, 2021) that examine the effect of using additional matching variables on DIF identification. While some items examined in the study were labeled as DIF in univariate DIF analyses, it was concluded that these items did

not show DIF when additional matching variables were used apart from the total score. Considering that the identified matching variables explain the DIF displayed by these items, it may be recommended to conduct a DIF analysis based on multivariate matching to control the first type of error in DIF studies.

It was determined that carrying away DIF items from the total score caused a variation in the number of items labeled as DIF although it did not yield consistent results. In other words, it can be argued that removing DIF items from the total score does not yield consistent results. This finding is parallel with the findings of studies (French & Maller, 2007; Lee & Geisinger, 2016; Svetina & Rutkowski, 2014). It was revealed that the exclusion of DIF items (removal of DIF items) while calculating the total score, which is the matching variable, affects the DIF detection power of the DIF detection technique. To eliminate the error caused by including DIF items in the total score calculation in DIF studies, and to balance Type I error and test power, it is considered appropriate to exclude DIF items from the total score.

One of the most basic assumptions of international assessment studies is that tests are equivalent in all languages or cultures. However, even in DIF analysis based on multivariate matching, some items were found to have displayed DIF. Considering that the poor quality of the translation makes the validity of the test scores, and therefore the comparability and interpretation of the scores impossible (Gierl, 2000), it is thought that translations in cross-cultural assessment studies should be done with an adaptation approach. However, since the selection of reading texts is of great importance in cross-cultural assessment studies (Grisay, Gonzalez & Monseur, 2009), the selection of these texts can be given particular importance.

In this study, multivariate DIF studies through logistic regression were performed. In a future study, a multivariate DIF analysis could be conducted based on IRT. Additionally, the removal of DIF items from the total score in this study was performed using logistic regression. A similar DIF study could also employ the Mantel-Haenszel method or another suitable DIF detection technique. Furthermore, this study utilized a literature review and multiple linear regression analysis to identify additional matching variables. In future research, alternative statistical methods, such as multilevel modeling, or judgmental approaches could be used to identify extra matching variables.

Acknowledgments

This study is a part of the doctoral dissertation of the first author under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Ahmet Yıldırım: Literature Review, Resources, Methodology, Data Analysis, Reporting and Writing-original draft. **Nizamettin Koç:** Supervision. Authors may edit this part based on their case.

Orcid

Ahmet Yıldırım https://orcid.org/0000-0002-0856-9678
Nizamettin Koç https://orcid.org/0000-0001-5412-0727

REFERENCES

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC.

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

Allalouf, A., & Sireci, S.G. (1998, April). *Detecting sources of DIF in translated verbal items* [Paper presentation]. American Educational Research Association 1998. San-Diego.

Arıkan, S., Van de Vijver, F.J.R., & Kutlay, Y. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education and Science*, *4*(1), 69-81.

Arffman, I. (2010, August). *Identifying translation-related sources of differential item functioning in international reading literacy assessments* [Paper presentation]. European Conference on Educational Research 2017. Helsinki.

Boughton, K.A., Gierl, M.J., & Khaliq, S.N. (2000, May). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance* [Paper presentation]. Canadian Society for Studies in Education. Alberta.

Braun, M., & Harkness, J.A. (2005). Text and context: Challenges to comparability in survey questions. Zlotnik, J.H.P. & Harkness, J. (Eds.). *Methodological aspects in cross-national research* (pp. 95-107). Mannheim: Zuma.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129-147.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Crane, P.K., Gibbons, L.E., Jolley, L., & Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care*, *44*(11), 115-123.

Çet, S. (2006). *A multivariate analysis in detecting differentially functioning items through the use of programme for international student assessment (PISA) 2003 mathematics literacy items* [Unpublished doctoral dissertation, Orta Doğu Teknik Üniversitesi]. Ankara.

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

Ercikan, K., Gierl, M.J., McCreith, T., Gautam, P., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*(3), 301-321.

French, B.F., & Maller, S.J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373-393.

Furlow, C.F., Ross, T.R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*(6), 441-464.

Gierl, M.J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, *25*(4), 280-296.

Gierl, M.J. (2004, April). *Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences* [Paper presentation]. American Educational Research Association. San Diego.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000, April). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large* [Paper presentation]. American Educational Research Association 2000. New Orleans.

Gierl, M.J., & Khaliq, S.N. (2000, April). *Identifying sources of differential item functioning on translated achievement tests: A confirmatory analysis* [Paper presentation]. National Council on Measurement in Education 2000. Louisiana, New Orleans.

Gierl, M.J., Rogers, W.T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF* [Paper presentation]. National Council on Measurement in Education 1999. Montréal, Quebec.

Gradshtein, M.F., Mead, A.D. & Gibby, R.E. (2010). *Making cognitive ability selection tests indifferent across cultures: The role of translation vs. national culture in measurement equivalence*. Retrieved October 20, 2015, from http://mypages.iit.edu/~mead/GradshteinMeadGibby-2010-10-01.pdf

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 63-83.

Hambleton, R.K. (1993). *Translating achievement tests for use in cross-national studies*. Retrieved December 21, 2016, from http://files.eric.ed.gov/fulltext/ED358128.pdf

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). *Advances in the detection of differentially functioning test items*. Retrieved October 20, 2016, from http://files.eric.ed.gov/fulltext/ED356264.pdf

Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*(1), 1-30.

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

Hidalgo, M.D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

International Association for the Evaluation of Educational Achievement. (2012). *TIMSS 2011 international results in mathematics*. Lynch School of Education, Boston College.

Jodoin, M.G. (1999). *Reducing Type I error rates using an effect size measure with the logistic regression procedure for DIF detection* [Unpublished Master's Thesis, University of Alberta]. Alberta.

Karasar, N. (2011). *Bilimsel araştırma yöntemleri* [*Scientific research methods*]. Ankara: Nobel Publishing.

Khalid, M.N., & Glas, C.A.W. (2013). A step-wise method for evaluation of differential item functioning. *Journal of Quantitative Methods*, *8*(2), 25-47.

Lee, H., & Geisinger, K.F. (2016). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, *76*(1), 141-163.

Mcdonald, R.P., & Ringo Ho, M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2015). *International large-scale assessments: Origins, growth and why countries participate in PISA*. OECD Publishing.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, *6*(2), 1-3.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355-371.

Sireci, S.G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*(2), 148-166.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2-3), 170-187.

Sireci, S.G., & Swaminathan, H. (1996). Evaluating translation equivalence: So what's the big DIF? Retrieved October 19, 2015, from http://files.eric.ed.gov/fulltext/ED428119.pdf

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. Retrieved October 20, 2015, from http://www.largescaleassessmentsineducation.com/content/pdf/s40536-014-0004-5.pdf

Tabachnick, B., & Fidell, L. (2013). Using multivariate statistics (5th Edition). Allyn & Bacon/Pearson Education.

Van de Vijver, F., & Tanzer, N.K. (2004). *Bias and equivalence in cross-cultural assessment: An overview*. Retrieved October 21, 2015, from http://resilienceresearch.org/files/article-vandevijver_tanzer.pdf

Wen, Y. (2014). *DIF analyses in multilevel data: Identification and effects on ability estimates* [Unpublished doctoral dissertation, University of Wisconsin-Milwaukee]. Wisconsin.

Yıldırım, H.H., & Yıldırım, S. (2011). Correlates of communalities as matching variables in differential item functioning analyses. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 40*, 386-396.

Yılmaz, M. (2021). Eğilim puanları kullanılarak abide çalışmasındaki maddelerin değişen madde fonksiyonu açısından incelenmesi [Unpublished Master's Thesis, Hacettepe University]. Ankara.

Zheng, Y., Gierl, M.J., & Cui, Y. (2007). *Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and logistic regression procedures* [Paper presentation]. National Council on Measurement in Education 2007. Chicago.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistics regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.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.