PISA 2015 Reading Test Item Parameters Across Language Groups: A measurement Invariance Study with Binary Variables*

Pelin BAĞDU SÖYLER ** Burak AYDIN *** Hakan ATILGAN ****

Abstract
Large-scale international assessments, including PISA, might be useful for countries to receive feedback on their education systems. Measurement invariance studies are one of the active research areas for these assessments, especially cross-cultural and linguistic comparability have attracted attention. PISA questions are prepared in the English language, and students from many countries answer the translated form. In this respect, the purpose of our study is to investigate whether there is a measurement invariance problem across native English and non-native English speaker groups in the PISA-2015 reading skills subtest. The study sample included students from Canada, the USA, and the UK as the native speaker group and students from Japan, Thailand, and Turkey as the non-native speaker group. Measurement invariance studies taking into account the binary structure of the data set for these two groups revealed that eight of the twenty-eight items in the PISA-2015 reading skills test had possible limitations in equivalence.

Key Words: PISA 2015, measurement invariance (MI), binary variables, reading skills.

INTRODUCTION
Internationally conducted student assessments play an essential role in the educational policies of countries. One of these assessments is administered by the OECD (Organization for Economic Cooperation and Development) (Milli Eğitim Bakanlığı-MEB, 2016). The OECD is an institution that plays a vital role in the regulation of the welfare of the world, economic development, and educational policies; it carries out many studies in line with its goals. One of these studies is the International Student Assessment Program (PISA), which is one of the most extensive educational researches in the world implemented internationally. PISA assessments are carried out regularly in fields of mathematics, science, and reading skills. In PISA, the concept of literacy is handled as special equipment used to fulfill a function in life practices. In this extensive study at the international level, equivalence studies are extremely important for ensuring the validity of the measurement instrument. PISA develops different cognitive measurement instruments to measure student performance at all levels in the fields of science and mathematics and contextual measurement instruments (OECD, 2018). One of the main assumptions in this practice, which closely concerns educational policies by comparing student achievements between countries, is that the measured structures are the same for all participants. Construct validity should be ensured by minimizing bias to make valid comparisons between different language groups and countries. Martin, Mullis, Gonzales, Gregory, Garden, O’Connor, Chrostowski and Smith (2000) emphasize the necessity of neutrality while comparing student achievement among countries. Accordingly, construct validity has distinctive importance.

*This work is based on the first author’s master thesis and preliminary results were presented at AERA 2021 Conference.
**Ministry of Education, İzmir-Türkiye, pelinbagdu@gmail.com, ORCID ID: 0000-0001-8169-2165
***Assoc. Prof., Ege University, Faculty of Education, İzmir-Türkiye, burak.aydin@ege.edu.tr, ORCID ID: 0000-0003-4462-1784
**** Prof., Ege University, Faculty of Education, İzmir-Türkiye, hakan.atilgan@ege.edu.tr ve 0000-0002-5562-3446, ORCID ID: 0000-0002-5562-3446

To cite this article:
Bağdu Söyler, P., Aydın, B., & Atılgan, H. (2021). PISA 2015 reading test item parameters across language groups: A measurement invariance study with binary variables. Journal of Measurement and Evaluation in Education and Psychology, 12(2), 112-128. doi: 10.21031/epod.800697

Received: 29.09.2020
Accepted: 09.06.2021
Baykal and Circi (2010) conducted a material revision study to improve the structure validity of PISA 2006 in science testing, and the authors concluded that the different characteristics of the countries should be taken into consideration in stages of item development and translation into different languages by examining the construct validity. Accordingly, it was seen that in international applications such as PISA, the tests are not understood by all participating countries in the same way. Generally, the active role of PISA in national education policies is based on the general assumption that PISA tests are reliable and valid instruments; therefore, this acceptance provides an international comparison of student performances. Researches on this have shown that there are many factors such as translation, item content, curriculum differences, exam motivation or exam anxiety, writing system, and culture. Linguistic diversity affects the comparability of scores and consequently may limit the validity of these studies (Arffman, 2002; Bonnet, 2002; Grisay & Monsseur, 2007; Hambleton, Merenda & Spielberger, 2005; He & van de Vijver, 2012; Kreiner & Christensen, 2014). PISA questions are prepared in English and are used by translating the languages of the countries whose native language is not English. The native language of most of the participating countries is not English, so non-native English-speaking countries use the tests translated into their language. Since PISA significantly affects the educational policies of countries, it is extremely important that the psychometric structure measured between countries and different groups is comparable (Brown, 2006). Scalar equivalence is required to compare the scores obtained from different language versions of the tests in a significant and valid way (Ercikan & Lyons-Thomas, 2013). In order to compare individuals from different cultures and languages in different subject areas, especially in a direct language-dependent area such as reading skills, it is a critical issue to have no equivalence problems in the structures measured by the tests and to ensure the measurement invariance of the tests.

Arffman (2010) identified six types of problems that limit the equivalence of PISA reading tests. These were language-specific differences in grammar, language-specific differences in writing, language-specific differences in meaning, differences in culture, translators’ choices and strategies, and problems with editing. Accordingly, it is important that the questions are accessible in terms of examining the factors that limit the equivalence of the items and understanding these problems. Based on the analysis of PISA 2006 reading items, Kreiner and Christensen (2014) pointed out that the validity of the measurement model was inadequate due to items with differential item functioning (DIF). As a result, it was not appropriate for countries to compare as such. Some critics have suggested that the PISA reading texts, to some extent, support Western countries, consistent with previous cultural and linguistic concerns. (Grisay et al., 2007; Grisay, Gonzalez & Monsseur, 2009; Oliveri & von Davier, 2011). Since countries with similar linguistic and cultural histories are likely to hold the equivalence in scores, it is predicted that the MI may be a problem for PISA assessments. (Asil & Gelbal, 2012; Kankaras & Moors, 2013).

In the literature, there are many MI studies in PISA student surveys. Asil and Gelbal (2012) investigated MI in terms of culture and linguistics in PISA 2006 student survey. Results revealed that as the cultural and linguistic differences between countries increase, the number of DIF items increases. Segeritz and Pant (2013) studied the Learning Approaches of Students (SAL) scale in the PISA 2003 in Germany sample among ethnic-cultural groups in a country. The findings obtained with the results have shown that the factor structure of the scale Learning Approaches between Germany and two immigrant student groups is comparable.

The equivalence of PISA tests between countries in terms of cultures and language is questionable. The main criticisms point to linguistic and cultural bias, potentially affecting the nature of reading tests. Therefore, the comparisons between countries raise doubts about accuracy. Literacy performance is influenced by a set of characteristics such as the nature of each language, the writing system used to stimulate literacy, the cultural style, teaching and learning approaches, and level of investment in socio-economic development and education (Asil & Brown, 2015).

MI of the cognitive data has been tested, and the cultural comparability correlations of the cognitive data have been examined by taking the technical reports as reference. It was concluded that comparing the total scores across different cultures may lead to incorrect results.
International large-scale applications such as PISA, TIMMS, and PIRLS aim to measure latent structures among participants and compare between groups. However, when these assessments participating in many countries are taken into consideration, some evidence has been obtained that the method is not practical in such large-scale assessments (Rutkowski & Stevina 2013; Ogretmen, 2006). Rutkowski and Stevina (2013) conducted a simulation study to investigate the change depending on the sample size and the number of groups of multi-group confirmatory factor analysis (MG-CFA) performance. In order to mimic real data, the data were simulated ordinal categorical and analyzed with a linear model. In the findings obtained, it was concluded that there is an inconsistent relationship between a sequential categorical data set and the linear model, so this method selection is not an excellent theoretical practice. In the findings obtained, it was concluded that there is an inconsistent relationship between an ordinal categorical data set and the linear model, so this method is not the right choice in theory. Readers are referred to Jöreskog, Sörbom, Toit and Toit (2001), Sirganci, Uyumaz and Yandi (2020), Gregoric, (2006), Salzberg et al. (1999), Önen (2009), Wu, Li & Zumbo (2007), and van de Schoot et al. (2013) for further reading on MG-CFA. Therefore, there is an operational need for the suitability of comparisons across countries. In PISA 2015, a recent approach has been applied for MI testing using item response theory (IRT) item consistency (OECD, 2016). Thus, the question raised about the reproducibility of these findings in the context of more common analysis techniques.

In order to compare individuals from different cultures with international measurement instruments, it is essential to hold the equivalence of their forms in different languages when the measurement instrument is translated into other languages. Therefore, measurement invariance is one of the most needed studies in cross-cultural comparisons of multiple groups. It is one of the preconditions to make correct decisions in terms of language skills of cultures and cross-language equivalence in a study playing a significant role in the educational policies of countries such as PISA. Thus, the construction validity studies are very important for the evidence of the validity of the measurement instrument. There are several studies in the literature regarding the MI of PISA; however, it is remarkable that many of the MI analyses ignore the binary nature of the PISA's data sets. PISA questions consist of multiple-choice and partial answer items. In assessments involving such items, it is crucial to perform the MI studies carefully using an appropriate method for the binary nature of the data set in order to achieve valid results.

**Measurement Invariance with Different Variable Types**

MI studies provide evidence of the structural validity of the measuring instrument. The equivalence of the characteristic of a psychological measurement instrument, such as construct validity and reliability, in different groups is defined as the measurement invariance (Herdman, 1998). Whether the psychological structure to be measured is comparable between groups in terms of different cultural factors or variables is essential for the validity. MI means that a measurement model has the same structure in multiple groups, and the factor structures and error variances of the items in the scale are equivalent (Bollen, 1989).

Evaluation of MI within common factor linear models is known as factorial invariance. When the linear factorial model is used in data sets involving binary, ordered, and Likert-type variables, the structure of the observed variables are ignored (Elosua, 2011). In order to test the MI, the chi-square difference test is used. However, the models are different for continuous and ordinal categorical datasets, so testing the MI between groups requires testing the parameters for each model (Meredith, 1993). While the related parameters are factor loadings and residual variance in a dataset containing continuous variables, the thresholds are required to compare between groups in an ordinal categorical dataset. Using the maximum likelihood estimation (ML) and continuous linear models to analyze ordered categorical datasets involves some disadvantages and uncertainties about the resource of invariance (Lubke & Muthén, 2004). French and Finch (2006) concluded that the chi-square difference test in evaluating measurement invariance was inadequate in a dataset containing multidimensional binary categorical items. Instead of the linear factor analysis commonly used for continuous variables, the variables in the ordered categorical structure can be modeled with MG-CFA in accordance with the threshold structure (Kim & Yoon, 2011). Since linear CFA is not a suitable analysis for ordered categorical data, the MI test cannot
be sufficiently compared with linear CFA (McDonald, 1999; Oishi, 2006; Reise et al., 1993). Meade and Lautenschlager (2004) stated that in some cases, the IRT approach could give different and potentially more useful information for modeling MI.

Without modeling the threshold structure, CFA assumes that the underlying distributions of dichotomous or polytomous variables are normal. Threshold values are mathematically related to item difficulty parameters in IRT (Lord & Novick, 1968; Takane & de Leeuw, 1987). Accordingly, ordered categorical CFA with the appropriate analysis method based on IRT to test the MI with ordered categorical variables gives more accurate results than linear CFA without considering the threshold structure (Kim & Yoon, 2011). It should be noted that, especially in PISA assessments, cognitive tests have a binary categorical structure, and attitude scales include Likert-type variables. In other words, analyzing categorical data using methods developed for continuous variables has serious limitations in general (Raykov, Marcoulides & Milsap, 2013).

**Measurement Invariance with Binary Variables**

It has been demonstrated in recent studies that the methods commonly used in MI studies have limitations. As mentioned previously, the MG-CFA method is frequently used for continuous, and Likert-type scored variables. Raykov, Dimitrov, Li, Marcoulides & Menold (2018) suggested an alternative method for testing the MI with binary scored items. This method aims to determine cases that do not hold the MI with item factor loadings and threshold values. The recent approach does not require defining a reference variable and allows us to study the MI directly with one or two-parameter IRT modeling (Raykov et al., 2018).

IRT suggests that the performance of a person in a test can be predicted according to the item characteristic curve that shows the relationship between the latent traits or abilities (Hambelton and Swaminathan, 1985). IRT is concerned with the participants’ responses to each item rather than the total score received from the test. Two item parameters can be used to define the item characteristic curve, which is the basis of IRT. One of these is item difficulty (b), and the other is item discrimination (a) index. Item difficulty states where the item is functional. For example, while an easy item is more functional for individuals with lower ability, a difficult item is more functional for individuals with higher ability levels. The item discrimination index states how well it characterized individuals who are below the ability level of the item and individuals with an ability level above this point (Baker, 2016).

Assume \( y = (y_1, y_2, ..., y_l) \) represents the components of a psychological scale. In addition, it is assumed that the component \( y \) discharges the conditions of structural invariance in groups with large samples (Millsap, 2011). In this setting, a factor analysis model has been developed in each group in which \( a \) parameter with loadings and \( b \) parameter with thresholds are related. Hence, the necessary conditions for \( y \) component and MI of the \( g^{th} \) group are represented as follows:

\[
\begin{align*}
    y_g^* &= \Lambda_g \eta_d + \delta_g \quad (1) \\
    \Lambda_{1} &= \Lambda_{2} = ... = \Lambda_g \\
    \tau_{1g} &= \tau_{2g} = ... = \tau_g \quad (2)
\end{align*}
\]

The pair of Equations 2 and 3 also represents a necessary condition to study a two-parameter IRT model or the DIF, a special case of it (Muthén, Asparouhov & Morin, 2015). DIF states that the probability of responding to the test item correctly is not an equality case in individuals with the same ability level and from different groups (Adams & Rowe, 1988). DIF analysis aims to investigate whether test scores are affected by variations from different groups and whether these variations give rise to a bias for any subgroup (Algina & Crocker, 1986). If the attribute measured by the test is the same in different subgroups, it can be seen that the items are affected by the same variability and that individuals with the same ability level are similar in the measured structure (Algina & Crocker, 1986). The MI analysis method in the binary scored items used in our study provided to test the MI by determining the items under the two-parameter IRT.
**Purpose of the Study**

The purpose of this study is to examine whether the PISA 2015 reading skills subtest is equivalent in terms of language skills for countries with native English and non-native English speakers. In order for comparisons and assessments to be valid, equivalence across cultures and languages should hold. Scales developed in a particular culture and language reflect characteristics of that culture and language. Translating a measurement instrument does not warrant that these two scales are equivalent (Sireci & Berberoğlu, 2000). It should be noted that the measurement instrument to be translated or adapted to another language will differ from its original form. These differences should be ensured to be acceptable in terms of psychometric properties (Hambleton & De Jong, 2003). In such a study that plays an essential role in the educational policies of countries, the intercultural equivalence of the tests in terms of language skills is one of the preconditions for making the right decisions (Arffman, 2010; Baykal & Circi, 2010; He, Barrera-Pedemonte & Bucholz, 2018). In this respect, it is very important to investigate construct validity carefully for the proof of the validity of the measuring instrument. Hence in this study, whether the reading skills test of the PISA 2015 assessment has MI problem between the translated language form and the original one has analyzed by statistical analysis methods.

**METHOD**

Sample sizes of PISA 2015 participant countries included in our study are 14157 from the UK, 5712 from the USA, 20058 from Canada, 6647 from Japan, 8249 from Thailand, and 5895 from Turkey. In PISA, not all students take the same test, and test forms contain common questions as well as different questions (OECD, 2016). A total of 64171 students from selected countries participated in the study. In PISA 2015, 66 different forms were prepared for countries that received computer-based tests. In our study, data from the 41st form were used given that it was the most frequently used form for Canada, UK, the USA, Japan, Thailand, and Turkey. Reading skills achievement was measured in this form with 28 items. The frequencies of the participants who took the 41st form in the sample by country are reported in Table 1.

Table 1 shows that the country with the highest number of participants is Canada with 34.4%, and the USA has the lowest number of participants with 8.9%. The sample of the study consists of 1524 students taken the 41st form from six countries separated out of the countries participating in PISA 2015. The countries included in the research were selected from the countries participating in the PISA 2015 with a computer-based assessment. Therefore, 28 items with the most responded form number 41 selected among 66 different forms were included in the analysis. This form included open-ended and multiple-choice questions. According to the type of question, the items are coded with 0 refers to false responses, 1 refers to partially correct responses, and 2 refers to correct ones. Since the model did not converge with only two partially scored items, the partially correct scores were treated as correct, and items 5 and 6 are re-coded as 0 for incorrect and 1 for correct responses. In our study, the ratio of the missing value to the total sample size was only 6%, considered low (Kline, 2016, p.83) and hence ignorable (Akbaş & Tavşancıl, 2015; Cheema, 2012; Downey and King, 1998; Rubin, 1976; Enders, 2010), and it was decided to exclude the missing data from the analysis to ease the model convergence.

**Data Analysis**

A single factor model was tested using CFA for each group. The item parameters obtained with separate CFA were examined. The full measurement invariance approach allows the item factor loadings and threshold values between the comparative groups to be the same, and the approximately defined measurement invariance approach allows only small differences in the parameters in question between the compared groups (Kim, Cao, Wang, & Nguyen, 2017). Muthén and Asparouhov (2013) bring in the term of approximate measurement invariance as a stage of measurement invariance, in addition to full invariance and partial invariance, with recent studies (van de Schoot et al., 2013).

Findings obtained in this direction have been reported.

Table 1. Sample Sizes Based On Countries
The countries included in this study are separated into two groups as native English (UK, Canada, USA) and non-native English speakers (Japan, Thailand, and Turkey). MI for binary scored items was tested using the Mplus 8.0 (Muthen & Muthen, 2019). In this direction, item loadings and threshold parameters were free for each item in MI analysis. The difference in BIC values (ΔBIC) between the baseline model (M₀) and the free model in each model were studied. The smaller the BIC value, the better the model-data fit (Nylund, Asparouhov & Muthén (2007). The model with ΔBIC> 10 indicates a strong misfit of the model, and such values are considered a threat to MI (Frank J., Fabozzi & Wiley, 2014).

RESULTS

In the first step, CFA was completed in accordance with the nature of binary variables for each group, and the model fit was examined. The model data fit findings obtained with CFA are presented in Table 2.

Table 2. Confirmatory Factor Analysis Results of Reading Skills Test PISA 2015

| Group (Countries)               | Chi-Square value | n     | RMSEA | CFI  | TLI  |
|---------------------------------|------------------|-------|-------|------|------|
| Native English Speakers         | 409.58*          | 1044  | .03   | .96  | .97  |
| Non-Native English Speakers     | 243.86*          | 480   | .03   | .97  | .98  |

*p<.05

When the model fit indices in Table 2 are examined, it is seen that the chi-square value is significant in both groups (p < .05). Based on the RMSEA values, it can be understood that the model fits perfectly in both groups since it is .03 for both groups. Concerning CFI and TLI fit indices, it is seen that the CFI value for the native language group indicates a strong fit with .96 and the TLI value with .97. The CFI and TLI values for the non-native English also indicate a strong fit with .97 and .98. CFA results indicated that the one-factor structure of PISA 2015 Reading Skills Test holds for both groups separately. Item factor loadings, threshold values, a and b parameters obtained as a result of the CFA analysis are showed in Table 3.
The item factor loadings, threshold values, \( a \) and \( b \) parameters obtained from the CFA to examine whether the item parameters of each group differ or not are given in Table 3. It is observed that the 21\textsuperscript{st} item has the least factor loading in the group with native language English, whereas the group with non-English has the greatest factor loading. Accordingly, while factor loadings are expected to be approximately equal with each other for both groups, this case indicates that the item does not work in the same way for both groups. It is understood that the 9\textsuperscript{th} and 18\textsuperscript{th} items in the group with native
language is English are the ones with the greatest factor loadings. The 15th item is the item with the least factor load (.45) in the group with non-native English and is close to the factor loading (.50) given by the other group in the 15th item. When the factor loadings and the parameters a of the 12th item are compared, the item factor loading of the group with native English is .99, and the parameter a is .63, whereas the item factor loading of the group with non-native English is .51 and the parameter a is .38. These values are substantially different for the items that are expected to measure the equal characteristic.

When we viewed the item threshold values and b parameters, whereas the threshold value is -1.13 for the threshold of the second item in the group with native English, in the group with non-native English, it is -.35, and b parameter is -2.07 in the group with native English; the group with non-native English is -0.51. These values are different for an item that should measure the same characteristic in both groups. Similarly, when the parameters of item 23 are compared in both groups, it is understood that the group with native English is -1.77 and -0.48 in the other group. The CFA results performed separately for the two groups are visually examined. It is difficult to say that items 2, 4, 6, 8, 9, 12, 15, 18, 21, 22, 23, 25, 26, 27, and 28 work similarly in psychometric terms. In order to examine whether the 15 differences determined visually are statistically significant, the variation of item parameters and BIC values in 56 different models were examined for the data set consisting of 28 items. The results of MI analysis in binary scored items for the groups with native and non-native English speakers are presented in Tables 4 and 5.

BIC values obtained from 56 different models to be free of item factor loading and thresholds for each item, their differences from the BIC value in the M_0 (ΔBIC) and item factor loadings and thresholds are given in Tables 4 and 5. The BIC values of the M_0 and the BIC values of each model were compared separately. The BIC value was found to be 44745.34 in M_0. The difference of BIC value in each model with BIC value of M_0 was calculated.
Table 4. Measurement Invariance Analysis of PISA 2015 Reading Skills Test Thresholds

| Model | Par | BIC     | ΔBIC | Group 1 | Group 2 |
|-------|-----|---------|------|---------|---------|
|       |     |         |      | λ       | t       | λ     | t     |
| M1    | t1  | 44748.62| 3.28 | -       | -1.69   | -     | -1.27 |
| M2    | t2  | 44708.93| -36.41*| -       | -2.35   | -     | -1.25 |
| M3    | t3  | 44737.70| 7.64 | -       | -2.69   | -     | -2.02 |
| M4    | t4  | 44748.02| 2.68 | -       | -2.21   | -     | -1.87 |
| M5    | t5  | 44746.98| 1.64 | -       | -1.84   | -     | -1.48 |
| M6    | t6  | 44752.57| 7.23 | -       | -2.14   | -     | -2.19 |
| M7    | t7  | 44739.41| -5.93| -       | 1.66    | -     | 1.00  |
| M8    | t8  | 44751.38| 6.04 | -       | 0.71    | -     | 0.87  |
| M9    | t9  | 44752.39| 7.05 | -       | -2.76   | -     | -2.66 |
| M10   | t10 | 44752.62| 7.28 | -       | 1.65    | -     | 1.61  |
| M11   | t11 | 44751.68| 6.34 | -       | -0.10   | -     | -0.24 |
| M12   | t12 | 44731.95| -13.39*| -       | -0.14   | -     | -0.72 |
| M13   | t13 | 44749.29| 3.95 | -       | -1.78   | -     | -2.10 |
| M14   | t14 | 44748.84| 3.50 | -       | -0.10   | -     | -0.34 |
| M15   | t15 | 44752.65| 7.31 | -       | 0.64    | -     | 0.65  |
| M16   | t16 | 44748.44| 3.10 | -       | -0.42   | -     | -0.17 |
| M17   | t17 | 44749.87| 4.43 | -       | -1.79   | -     | -2.05 |
| M18   | t18 | 44745.93| 0.59 | -       | -2.01   | -     | -2.44 |
| M19   | t19 | 44751.14| 5.80 | -       | -0.13   | -     | -0.29 |
| M20   | t20 | 44742.71| -2.63| -       | -0.84   | -     | -0.37 |
| M21   | t21 | 44751.50| 6.16 | -       | -0.21   | -     | -0.71 |
| M22   | t22 | 44691.77| -53.57*| -       | 1.13    | -     | -1.58 |
| M23   | t23 | 44745.36| 0.02 | -       | -1.66   | -     | -1.30 |
| M24   | t24 | 44752.56| 7.22 | -       | 0.31    | -     | 0.27  |
| M25   | t25 | 44722.68| -22.66*| -       | -1.15   | -     | 0.58  |
| M26   | t26 | 44725.27| -20.07*| -       | -1.75   | -     | -2.66 |
| M27   | t27 | 44743.97| -1.37| -       | -0.59   | -     | -0.97 |
| M28   | t28 | 44743.49| -1.85| -       | 1.32    | -     | 0.82  |

Note: λ= item factor loadings; t=threshold; Grup 1: Native English Speakers  Grup 2: Non-Native English Speakers
Table 5. Measurement Invariance Analysis of PISA 2015 Reading Skills Test Item Factor Loadings

| Model | Par | BIC     | ΔBIC | ʎ  | t     | ʎ  | t     |
|-------|-----|---------|------|-----|-------|-----|-------|
| M29   | ʎ₁  | 44747.36| 2.02 | 1.15|       | 1.57|       |
| M30   | ʎ₂  | 44722.44| -22.90*| 1.09|       | 2.09|       |
| M31   | ʎ₃  | 44741.84| -3.50 | 1.20|       | 1.73|       |
| M32   | ʎ₄  | 44752.67| 7.43  | 1.19|       | 1.20|       |
| M33   | ʎ₅  | 44751.06| 5.72  | 1.36|       | 1.17|       |
| M34   | ʎ₆  | 44752.55| 7.21  | 1.68|       | 1.74|       |
| M35   | ʎ₇  | 44749.53| 4.19  | 1.30|       | 1.74|       |
| M36   | ʎ₈  | 44752.60| 7.26  | 1.09|       | 1.14|       |
| M37   | ʎ₉  | 44752.26| 6.92  | 1.75|       | 1.65|       |
| M38   | ʎ₁₀ | 44746.56| 1.12  | 0.92|       | 1.49|       |
| M39   | ʎ₁₁ | 44748.68| 4.34  | 1.40|       | 1.05|       |
| M40   | ʎ₁₂ | 44731.21| -24.13*| 1.06|       | 0.44|       |
| M41   | ʎ₁₃ | 44751.05| 5.71  | 1.52|       | 1.33|       |
| M42   | ʎ₁₄ | 44752.08| 6.74  | 0.89|       | 0.78|       |
| M43   | ʎ₁₅ | 44752.64| 7.30  | 0.50|       | 0.51|       |
| M44   | ʎ₁₆ | 44750.76| 5.42  | 0.59|       | 0.77|       |
| M45   | ʎ₁₇ | 44749.77| 4.43  | 1.28|       | 1.05|       |
| M46   | ʎ₁₈ | 44736.14| -9.20*| 1.75|       | 1.15|       |
| M47   | ʎ₁₉ | 44751.27| 5.93  | 0.94|       | 0.77|       |
| M48   | ʎ₂₀ | 44749.49| 4.15  | 1.46|       | 0.86|       |
| M49   | ʎ₂₁ | 44751.57| 6.23  | 1.46|       | 1.69|       |
| M50   | ʎ₂₂ | 44736.79| -8.55*| 1.18|       | 0.63|       |
| M51   | ʎ₂₃ | 44734.46| -10.88*| 0.97|       | 1.62|       |
| M52   | ʎ₂₄ | 44748.02| 2.68  | 0.75|       | 1.11|       |
| M53   | ʎ₂₅ | 44747.32| 1.98  | 0.78|       | 1.17|       |
| M54   | ʎ₂₆ | 44739.60| -5.71 | 0.95|       | 0.44|       |
| M55   | ʎ₂₇ | 44736.40| -8.64*| 1.12|       | 0.58|       |
| M56   | ʎ₂₈ | 44752.64| 7.30  | 1.43|       | 1.40|       |

Note: ʎ= item factor loadings; t=threshold; Grup 1: Native English Speakers  Grup 2: Non-Native English Speakers
Findings showed that ΔBIC value of the second item is -36.41 (ΔBIC > 10) in Model 2 and ΔBIC value is -22.90 (ΔBIC > 10) in Model 30. It is evaluated that the threshold values of the second item are quite different from each other, as -2.35 for the group (Group 1) with native English speakers and -1.25 for the group with non-native English speakers (Group 2). Accordingly, it can be said that the second item does not show the model fit and is not comparable for both groups. It is evaluated that ΔBIC value of item 18 in Model 46 is a poor fit with -9.20 (6 < ΔBIC < 10). Item thresholds and a, b parameters have different values from each other, as seen in Table 3. Similarly, the ΔBIC value of item 22 in Model 22 is -53.57 (ΔBIC > 10), and in Model 50 this value is -8.55 (6 < ΔBIC < 10). Table 3 is indicated that the parameters of these items differ from each other on the basis of both groups. Items 12, 23, 25, 26, and 27 also seem to have poor model fit. Therefore, it is evaluated that ΔBIC values of 8 in 28 items are not in the range of acceptable model fit, and item parameters differ parallel with these results.

DISCUSSION and CONCLUSION

In this study, the MI of the PISA 2015 Reading Skills Test in terms of the language variable between the countries with native English speakers and the countries with non-native English speakers was tested with binary scored items. For two groups with native and non-native English speakers, CFA was performed separately, and model fit was examined, and it was concluded that overall factor structures were confirmed for each group. Item parameters were compared in both groups with the findings obtained with CFA. It was understood that the factor loadings and threshold parameters of some of the items assumed to measure the same ability in both groups of the PISA 2015 Reading Skills test differ considerably from each other. Therefore, it was concluded that there could be a limitation for the comparability of the groups.

When the item thresholds and factor loadings of these items were compared, it was observed that there was a substantial difference. It was evaluated that 8 out of 28 items in the 41st form of PISA 2015 Reading Skills possibly limit the scalar equivalence. Such a limitation in at least one item means that the MI cannot be fully supported for the whole test (Raykov et al., 2018). Therefore, in this test, it can be concluded that the MI cannot be fully defensible without identifying sources that limit the comparison between English and non-native English groups. In the literature, there are similar MI findings. For example, Baykal and Circi (2010) studied the 2006 PISA science test. The authors asked teachers to evaluate the positive and negative properties of the items, an item evaluation form was created, and the items were categorized according to their content. Negative categories were determined according to culture-specific factors reflected in language, grammatical difficulties, unknown words, and expressions of sentences. Item revisions are completed based on the negative categories. A revised test was created by selecting 22 items from the Turkish version of the science test. With the revised science test, the original science test versions were administered to each of two equivalent groups consisting of 30 students. It was concluded that the group that took the language-wise revised test performed better in all the items compared to the group that took the original translation. A similar study by Asil and Brown (2015) compared the English version of the test and its versions translated into other languages of the PISA 2009 reading skills test. The authors reported that socio-economic factors significantly affect the MI, and linguistic factors are relatively less effective.

In international assessments such as PISA, the questions prepared in English are translated into another language by the expert translators and then translated back to English to ensure its equivalence with the original version. In order to study these factors carefully, information about the effects of the differences in culture and their reflections in the language should be obtained in measurement instruments (Goldstein, 2017). Items that are specific to a language and contain expressions causing bias should be excluded from the test. PISA 2015 science test items are not publicly available, the items that limited the MI could not be examined, and the differences between the results could not be studied in detail.
REFERENCES

Adams, R., & Rowe , K. (1988). *Educational research, methodology, and measurement: An international handbook*. Oxford: Pergamon Press.

Akbaş, U. & Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 38-57.

Algina, J. & Crocker, L., (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Arffman, I. (2002). In search of equivalence: Translation problems in international literacy studies. Finland.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, 54(1), 37-59.

Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. *Education and Science*, 236-249.

Asil, M., & Brown, G. (2015). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, 16(1), 71-93.

Baker, F. B. (2016). *The basics of item response theory*. Ankara: Pegem Academy.

Baykal, A., & Circi, R. (2010). Item revision to improve construct validity: A study on released science items in Turkish PISA 2006. *Procedia Social and Behavioral Sciences*, 2(2), 1931-1935.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, 9(3), 387-399.

Brown , T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Cheema, J. (2012). Handling missing data in educational research using SPSS. Unpublished doctoral dissertation. *George Mason University*.

Downey, R., & King, C. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, 175-191.

Elosua, P. (2011). Assessing Measurement Equivalence in Ordered-Categorical Data. *Psicológica*, 403-421.

Enders, C. K. (2010). *Applied missing data analysis*. (1. Ed.). New York: The Guilford Publications, Inc.

Ercikan, K., & Swaminathan , H. (1985). *Analysis of General Psychology*.

Elosua, P. (2011). Assessing Measurement Equivalence in Ordered-Categorical Data. *Psicológica*, 403-421.

Ercikan, K., & Swaminathan , H. (1985). *Analysis of General Psychology*.

Grisay, A., de Jong, J. H., Gebhardt, E., Bereznier, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3), 249-266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *ERI Monograph Series: Issues and Methodologies In Large-Scale Assessments*, 2, 63-84.

Hambelton, R. K., & Swaminathan , H. (1985). *Item Response Theory*. Nijhoff Publishing.

Hambelton, R. K., & De Jong, J. A. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing*, 127-134.

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

Herdman M., Rushby J. F., & Badia X. (1998). A Model of Equivalence in The Cultural Adaptation of HRQol Instruments: The Universalist Approach. *Quality Of Life Research*, 7(4), 323-335.

He, J., Barrera-Pedemonte, F., & Bucholz, J. (2018). Cross-cultural comparability of non-cognitive constructs in TIMSS and PISA. *Assessment in Education:Principles, Policy & Practice*, 26(4), 369-385.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*.

Jöreskog, K.G., Sörbom, D., Du Toit, S.H.C., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd ed.). Lincolnwood, IL: Scientific Software International.

Kankaras, M., & Moors, G. (2013). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, 43(3), 381-399.
Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling*, 212-228.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford publications.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model ranking of countries according to reading literacy. *Psychometrika*, 210-231.

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

Lubke, G. H., & Muthen, B. O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(4), 514-534.

Martin, M., Mullis, I., Gonzalez, E., Gregory, K., Smith, T., Chrostowski, S., O'Connor, K. (2000). TIMSS 2009 International Science Report: Findings from IEA's Repeat of The Third International Mathematics and Science Study at the Eighth Grade.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361-388.

Milli Eğitim Bakanlığı (2016). *PISA 2015 International report*. Ankara. [Online: https://ods.gov.meb.gov.tr/www/2015-pisa-ulusalraporu/icerik/204], 2016.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*.

Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York: US: Routledge/Taylor & Francis Group.

Muthén, B., and Asparouhov, T. (2013). *BSEM measurement invariance analysis*. Mplus Web Notes: No. 17. Available online at: www.statmodel.com

Muthén, B., Asparouhov, T., & Morin, A. J. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal Of Management*.

Muthén, L. K., & Muthén, B. O. (2019). *Mplus user’s guide*. Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.

Organisation for Economic Co-operation and Development (2016). Online: http://www.oecd.org/education/

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, 40(4), 411-423.

Ogretmen, T. (2006). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi*: Türkiye-Amerika Birleşik Devletleri örneği. Ankara.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling*, 53(3), 315-333.

Onen, E. (2009). *Ölçme değişmezliğin yapısal eşitlik modellenmesi teknikleri ile incelenmesi*. Ankara: Ankara University, Doctoral Thesis.

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. *Educational and Psychological Measurement*, 73(4), 713-727.

Raykov, T., Dimitrov, D., Marcoulides, G., Li, T., & Menold, N. (2018). Examining Measurement Invariance and Differential Item Functioning With Discrete Latent Construct Indicators: A Note on a Multiple Testing Procedure. *Educational and Psychological Measurement*, 78(2), 343-352.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552-566.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Salk, T., Sinkovics, R., & Schlgelmich, B. (1999). Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches. *Australasian Marketing Journal*, 23-38.

Sirici, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, 13(3), 229-248.

Sirganci, G., Uyumaz, G., & Yandi, A. (2020). Measurement invariance testing with alignment method: Many groups comparison. *International Journal of Assessment Tools in Education*, 657-673.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.
van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 1–15.

Wu, D., Li, Z., & Zumbo, B. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), 1-26.
Appendix A. Mplus 8.0 Syntax for CFA

TITLE: CFA for the first group (native English)
DATA: FILE IS ING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;

TITLE: CFA for the second group (non-English)
DATA: FILE IS NONING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;
Appendix  B. Mplus 8.0 Syntax for the MI with Binary Variables

M₀ base model:
TITLE: Raykov (2018) M0
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;

Example syntax to relase a threshold (M₁-M₂₈):
TITLE: Raykov (2018) M1 (relase first threshold)
!LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u2$1-u28$1](T2-T28);
[u1$1*];
[f1*];
f1*;
Example syntax to release a loading(M29-M56):

TITLE: Raykov (2018) M29 (relase first loading)
LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;

MODEL:

%OVERALL%

f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1 @1;

%C#2%

f1 BY u1+
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;
PISA 2015 Dil Gruplarına Göre Madde Parametreleri:
İkili Değişkenlerle Ölçme Değişmezliği Çalışması *

Pelin BAĞDU SÖYLER ** * Burak AYDIN *** Hakan ATILGAN ****

Öz
PISA gibi uluslararası düzeyde yapılan sınavlarda, ülkelerin eğitim sistemlerinin etkililiğini hakkında değerlendirmeler yapılmaktadır. Dolayısıyla bu uygulamalar için hazırlanan Ölçme araçlarının geçerliği incelenirken farklı değişkenlerle göre değerlendirilirinin sınıması önemli konulardan biridir. PISA uygulamasının soruları İngilizce dilinde hazırlanmaktadır. Birçok ülkeden katılan öğrenciler, testin orijinal formunu değil çeviri formunu cevaplamaktadırlar. PISA’nın uygulama dilinden farklı bir dil kökeninde eğitim gören öğrencilerle testi orijinal formda alan öğrenciler arasında dil değişkenine göre bir ölçe değişmezliği sorunu oluştuğu görülmektedir. Bu bağlamda çalışmanın amacı PISA-2015 okuma becerileri alt testinde ana dili İngilizce olan ülkeler ile ana dili İngilizce olmayan ülkelerarasında çeviri kaynaklanan bir Ölçme değişmezliği sorunu oluşturmaktır. Araştırmanın amacı bir aracılığıyla ana dili İngilizce olan ülkelerin, ABD, Türkiye ve İngiltere’den katılan ülkelerden Japonya, Tayland ve Türkiye örneklerini araştırmaya dahil edilmiştir. Bu araştırında PISA-2015 okuma becerileri testinin ana dili değişkenine göre ölçe değişmezliği veri setinin ikili kategorik yapısına uygun olarak test edilmiştir. Yapılan analizler ile elde edilen bulgular doğrultusunda, PISA-2015 okuma becerileri testindeki 28 maddeden 8’inin ana dili İngilizce olmayan ülkelerin eğitiminin ikili değişkenlerle ölçülebileceğini öne sürülmiştir. PISA gibi uluslararası sınavların sonucu ülkeler arasında karşılaştırılabilir olması için Ölçme değişmezliğini sınırlandıran faktörlerin belirlenerek etkinin en aza indirilmesi gerektiği önerilmiştir.

Anahtar Kelimeler: PISA 2015, ölçe değişmezliği, ikili değişkenlerde çok gruplu ölçe değişmezliği analizi

GİRİŞ
Ulkerlerin eğitim politikalarında önemli rol oynayan uluslararası sınavlardan biri de OECD (Organization for Economic Cooperation and Development) tarafından yürütülmektedir (MEB, 2016). OECD dünya halklarının refahının, ekonomik kalkınmalarının ve eğitim politikalarının düzenlenmesinde önemli rol oynayan bir kuruluştur. Amaçları doğrultusunda pek çok çalışmalar yapılmaktadır. Bu çalışmaların bir de ülkemizde de katıldığı uluslararası boyutta uygulanın dünyının en büyük eğitim araştırımlarından olan Uluslararası Öğrenci Değerlendirme Programıdır (PISA-Programme for International Student Assessment). PISA uygulamaları; matematik, fen ve okuma becerileri konu alanlarında her üç yilda bir düzenlen olarak yapılmaktadır. Bu araştırımlar yapıldığında an kavram “okuryazarlık” üzerinde durulmaktadır. PISA da okuryazarlık kavramı, yaşam pratikleri içinde bir işlevi yerine getirme amaçla kullanılan bireysel bir donanım olarak ele alınır. Uluslararası düzeyde ülkelerin eğitim çekitlerinin değerlendirilmesi geniș çaplı araştırılarda, eğitim seviyelerinin karşılaştırılabilmesi için ölçe acısının geçerli olduğuna dair bulgular dolayısıyla değerlendirilme işleme çalışması son derece önemlidir.

Ulkerler arasında öğrenci başarlarını karşılaştırarak eğitim politikalarının yakından ilgilendiren PISA uygulamasında temel varsayımlardan biri de ölçülen becerilerin tüm katılımcılara aynı olmasıdır. Farklı dillerdeki ve ülkeler arasında geçerli karşılaştırma yapmak için yanılışı en aza indirerek yapın

* Bu çalışma “PISA 2015 Okuma Becerileri Testinin Ana Dili Değişkenine Göre Ölçe Değişmezliğinin İncelenmesi” isimli yüksek lisans tezinden üretilmiştir. AERA 2021’de yazılı bildiri olarak sunulmuştur.
** Milli Eğitim Bakanlığı, İzmir-Türkiye, pelinbagdu@gmail.com ve 0000-0001-8169-2165
*** Doç. Dr., Ege Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, burak.aydin@ege.edu.tr ve 0000-0003-4462-1784
**** Prof. Dr., Ege Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, hakan.atilgan@ege.edu.tr ve 0000-0002-5562-3446

Bu makaleye atıfta bulunmamak için:
Bağdu Söyler, P., Aydin, B., & Atilgan, H. (2021). PISA 2015 reading test item parameters across language groups: A measurement invariance study with binary variables. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 112-128. doi: 10.21031/epod.800697

Geliş Tarihi: 29.09.2020
Kabul Tarihi: 09.06.2021
Bu doğrultuda yapılan araştırmaların doğruluğu hakkında şüphe uyandırmaktadır. Genel olarak, PISA’nın ulusal eğitim politikalarında etkin rolü, PISA testlerinin güvendiği ve geçerli araçlar olduğu genel kabulüne dayanmaktadır; dolayısıyla bu kabul öğrenci performanslarının uluslararası düzeyde karşılaştırılması sağlanabilir. Bununla ilgili yapılan araştırmalar öncelikle ölçme değişmezliğini, ölçme değişmezliğinin sağlanması önemlidir. Bu doğrultuda, ölçümlerin karşılaştırılabilirliğini açığa çıkaran ve bu problemi ele alan birçok çalışma bulunmaktadır (Arffman, 2010; Bonnet, 2002; Grisay ve Monsieur, 2007; Hambleton, Merenda ve Spielberger, 2005; He ve van de Vijver, 2012; Kreiner ve Christensen, 2014).

Elde edilen bulgular sonucunda özelliklerin potansiyeli olarak değerlendirilmesi, ölçme değişmezliğini önemlendirir. Arffman (2010), PISA okuma metinlerinin eş değerliğini sınırlayan altı tip problem belirtmiştir. Bunlar dil, dilbilgisindeki farklılıklar, kültürdeki farklılıklar, çeviri,(mutlak) fonksiyon paylarının karşılaştırılabilirliğini etkiler ve sonuç olarak bu çalışmaların geçerliliğini sınırlayabilir (Arffman, 2002; Bonnet, 2002; Grisay ve Monsieur, 2007; Hambleton, Merenda ve Spielberger, 2005; He ve van de Vijver, 2012; Kreiner ve Christensen, 2014). PISA soruları İngilizce olarak hazırlanmıştır. Katılımcı ülkelerin bir çoğunun ana dili İngilizce olduğu anlamına gelir, bu durum dilbilgisindeki farklılıklar, kültürdeki farklılıklar, çeviri, mutlak fonksiyon paylarının karşılaştırılabilirliğini etkiler (Brown, 2006; Oliveri ve von Davier, 2011). Bu doğrultuda PISA gibi uluslararası uygulamalarda testlerin kültür ve dille bağlı etkenlerden bağımsız olarak, katılımcı tüm ülkeler tarafından aynı şekilde analısalınının karşılaştırılabilirliğini açığa çıkaran önemli bir husus olduğu anlaşılmaktadır.

Ölçme ve Değerlendirme Dergisi
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 - 6575

Bu doğrultuda yapılmış olan çalışmalarda ölçme değişmezliği ayrı edici bir önem sahiptir. Baykal ve Circi (2010), PISA 2006 uygulamasının fen testinde yapı değerlendirme gelistirilmesi için madde düzeyle çalışma yapmışlardır. Aşırı derecede seçilmiş madde sayısı, tarama maddesi ve materyallerin uygulanması, dil ve kültürle ilgili olan görüşleri etkilemektedir. Bu doğrultuda PISA gibi uluslararası uygulamalarda testlerin kültür ve dille bağlı etkenlerden bağımsız olarak, katılımcı tüm ülkeler tarafından aynı şekilde analısalınının karşılaştırılabilirliğini açığa çıkaran önemli bir husus olduğu anlaşılmaktadır.
Bağdul Söyler, P., Aydin, B., Atulgan, H. / PISA 2015 Dil Gruplarına Göre Madde Parametreleri: İkili Değişkenlerle Ölçme Değişmezliği Çalışması

dil becerileri açısından kültür ve diller arası eş değerinin sağlanmış olması, doğru kararlar alınmasını ön koşullarından önemlidir. Dolayısıyla yapılı geçerlilik çalışmalarıın titizlikle yapılmasının, ölçme aracının geçerliliğini dair kanıt elde edilmesini sağlamaktadır. Ayrıca, PISA uygulamalarının ölçme değişmezliğine ilişkin çeşitli çalışmaları mevcuttur. Yapılan çalışmaların birçoğunda seçilen Ölçme değişmezliği analiz yöntemlerinin veri setine uygunluğunun göz önünde bulundurulmaması dikkat çekicidir. Ölçme değişmezliğini çalışmalara güvenir sonuçlarla yapılmasını için seçilen yöntemin veri setinin hızla uygun olması mutlaka göz önünde bulundurulmalıdır.

Ölçme değişmezliği çalışmaları ölçme aracının yapı geçerliliğine dair kanıt elde edilmesini sağlar. Psikologî bir ölçme aracının yapı geçerliliğini ve güvenilirği gibi özelliklerinin farklı gruplardaki eşitliği Ölçme değişmezliği (esdeğerliği) olarak tanımlanır. Ölçme değişmezliği, bir ölçme modellinin çoklu gruplara uygun hale getirilmesi durumlarda önemlidir (Bollen, 1989).

Ölçme değişmezliğinin olma faktör doğrusal modelleri çerçevesinde değerlendirilmesi faktöriyel değişmezlik olarak bilinir. Bu metodoloji, bir ölçme modelinin farklı gruplar arasındaki parametrelerin esdegerliği değerlendirilmesini gerektirir. Doğrusal faktöriyel model iki, sıralı ve ikiert tipi değişkenler için veri setindeki kullanıldığında migli olan değişkenlerin yapı ile gözardı edilmiş olur (Elosua, 2011). Ölçme değişmezliğini test etmek için serbest modellerdeki k-tane fark testi yapıldığı gibi çok gruplu faktör analizleri de kullanılmaktadır. Faktör analizlerinde sıralı kategorik veri setleri için modeller farklıdır, bu nedenle gruplar arası ölçme değişmezliğinin analiz ve parametrelerin her bir model için test edilmesini gerektirir (Meredith, 1993). Sırali değişkenler için bir veri setinde iki bilinir parametreler faktör yüklerini ve artırmaya yana kılan farklı gruplar arasındaki farklılık yapabilecek modelin veri setindeki kullanılarak test edilmiş bir faktör analizi (ÇGDFA) ile modellenmelerdir (Kim ve Yoon, 2011). Ölçme değişmezliği testi, doğru ölçme değişmezliği testi, doğrusal doğrulayıcı faktör analizi (DFA) ile madde tepki kuramına (MTK) dayalı DMF analizi ile yeterince karşılaştırılmaz, çünkü doğrusal DFA sıralı kategorik veriler için uygun bir analiz değildir (McDonald, 1999; Oishi, 2006; Reise ve diğerleri, 1999). Meade ve Lautenschlager (2004), bazı durumlarda MTK yaklaşımının ölçme değişmezliğini test etmek için en büyük olabildirlik kestirimi (ML) ile sürekli doğrusal modellerin kullanılması bazı dezavantajlar ve değişmezliğin kaynağı hakkında belirsizlikler içermiştir (Lubke ve Mathén, 2004). French ve Finch (2006), artışın çözümlerinde çok boyutlu ikili kategorik modeller sonucunda ki-kare fark testinin ölçme değişmezliğini değerlendirmededeki gücüne oldukça düşük olduğu sonucuna varmışlardır. Sürekli değişkenler için yaygın olarak işe kulan doğrusal faktör analizi, bir ölçme ve değerleme (latent variable modelling) ve çoklu test metodu (multiple testing method) kullanılır. Birinci adımda ölçümlerin ölçme değişmezliğini incelemek için en sik kullanılan yöntemlerden biri ÇGDFA’dır (Jöreskog, Sörbom, Toit ve Toit, 2001; akt. Sirganci. Uyumaz ve Yandi, 2020). ÇGDFA aynı zamanda farklı gruplar arasında ölçümlerin veri setindeki uygunlukları için de karar verilebilir. ÇGDFA analizleri ve ölçme değişmezliği analizleri analiz edirken gruplardan biri referans grup olarak belirlenir ve bu grup hâlindeki ve diğer grupta sağlıklı verilerin doğrulanıp dökülmuş olup olmadığını ölçme değişmezliği olarak test edilmektedir. Her bir aşamada ayrıca seçilmiş olan koşul olarak kabul edilir (Wu, Li ve Zumbo, 2007). Sıralı olmayan
değişkenlerin (ör. sıra kategorik) normal dağılımlar bir örtük değişkenin yansıması olduğu kabulü ve bu kabule uygun tahminleyicilerle ÇG DFA tamamlanabilir. Fakat Raykov, Marcoulides ve Millsap (2013) örtük değişken modelleme yaklaşımlarını yanlıs sonuçlar verebileceğini ve çok test yaklaşımlarının daha az sayıda smırnlılıği olduğunu belirtmiştir.

İkili Kategorik Değişkenlerden Ölçme Değişmezliği

Raykov ve ark. (2018) çok test yaklaşımlarının ikili değişkenlerin yapısına uygun olarak kullanılabileceğini, 9 sorudan oluşan bir matematik yeteneğ testinin 771'i erkek, 744'ü kız olan 1515 kişilik bir örnekleme uygulamasında elde edilen verilerde uygulayarak açıklamışlardır. Her bir matde için eşik değerler ve faktör yükleri serbest bırakılarak temel modelle (M0) yapılan karşıştırmanın göre ölçme değişmezliği incelenmiştir. Bu yaklaştırmada MTK ile ortuşmaktadır. MTK, bireyin bir testte geçtiğini belirttiği performansın örtük özellikler veya yetenekleri ve bir maddedeki performansına bağlı olarak her özelliği arasındaki ilişkili taşıyan madde karakteristik eğrilerine göre kestirilebileceğini öne sürer (Hamberlon ve Swaminathan, 1985). MTK her bir maddede verilen cevabin doğru ya da yanlıştırı ile ilgilenir. Bireyin birer matdeye vermiş olduğu doğru cevaplar 1, yanlış cevaplar 0 olacak şekilde ikili madde (binary item) formunda puanlanır (Baker, 2016). MTK’nın temeli olan madde karakteristik egrisinin tanımlanması için iki madde parametreleri kullanılır. Bunlardan biri madde güçlüüğü (b) diğer ise madde ayrıntılı edilici (a) indirgelidir. Madde güçlüüğü, madreinin hangi noktada işlevsel olduğunu ifade eder. Örnek olarak; kolay bir madde daha düşük yetenek düzeyindeki bireyler için daha işlevsel iken, zor bir madde yüksek yetenek düzeyinde bulunan bireyler için daha işlevseldir. Madde ayrıntılı edilici indeksi ise, madreinin bulunduğu noktası altında bulunan bireyler için bu noktannın üzerinde yetenekin en yüksek sahip bireylerin ne kadar iyi ayrıntılı edebildiğini gösterir (Baker, 2016).

Bir psikometrik ölçeğin bileşenlerini y = (y₁, y₂, . . . , yₖ) temsil etmektedir. Ayrıca bileşenin her biri büyük önemlere sahip gruplarda yapısal değişmezlikinin koşullarını yerine getirdiği varsaymaktadır (Millsap, 2011). Bir psikometrik ölçek ölçümlerinde a (yanlış) ve b (sahip) parametreleri ile birlikte diğer veri olarak g-rubunun ölçümü ile ölçüm değişmezliği için gerekli şartlar aşağıda verilmiştir.

\[ y = \Lambda \xi + \delta \]

\[ \Lambda = \Lambda_1, \Lambda_2, \ldots, \Lambda_k \]

\[ \tau_1 = \tau_2 = \ldots = \tau_k \]

Eşitlik 2 ve 3 ayrıca iki parametreli MTK modeli ya da MTK’nin bir özel durumunu olan DMF incelemek için yeterli bir şartı temsil etmektedir (Musken, Asparouhov ve Morin, 2015). Bu yöntemde ilk aşama g-rubunun hâli almak üzere, kategorik veri setlerinde en büyük kabulü ve bu kabulü için yeterli şartı özelliklerini kullanılmaktadır (Musken ve Muthen, 2016). Çok gruplu bu modelde: (i) grup özemi için tüm faktör yükleri ve eşik değerleri sabit tutulmuştur, (ii) örtük değişkenlerin ortalamaları ve varyansları yalnızca birinci grupta serbest kalınmıştır, (iii) q>1 ise örtük kovaryans matrisinin köşegen olmayan öğeleri tüm gruplarda serbest bırakılır. Dolayısıyla birincı aşamada uygulanan 2 ve 3 numaralı denklem çiftleri, yalnızca birinci grupta tüm faktörlerin ortalamaları ve varyanslarını sırasıyla 1 ve 0 olarak belirlenmiştir, (iiii) q>1 ise örtük kovaryans matrisinin köşegen olmayan öğeleri tüm gruplarda serbest bırakılır. Dolayısıyla birincı aşamada uygulanan 2 ve 3 numaralı denklem çifti, yalnızca birinci grupta tüm faktörlerin ortalamaları ve varyanslarını sırasıyla 1 ve 0’a sabitlenip, kalınlarının tüm gruplarda serbest bırakılmasını gerektirir ve q=1 olması durumunda örtük kovaryans matrisinin köşegen olmayan öğeleri birinci (referans) grupta 1’e sabitlenmiş ana köşegen elemanları dışında tüm gruplarda serbest bırakılır. Bu sınırlılar darnıcık g-rubun modeli sürümlüğünü M₀ olarak gösterilmiştir. İçindeki aşamada bir maddein esik değerleri yanı sıra faktör yükleri de , M₀’a tekli parametreler olarak arda serbest bırakılır ve bu da 2k ilgili model M (1),…M(2k) modelini oluşturur (Raykov, Marcoulides ve Millsap, 2013). Bu 2k modellerin her birinde M₀’yuvananmış (nested) özelliği bulunmaktadır, yanı M₀ M(1’den M(2k)’ya kadar tüm modellerde grupparsal eşdeğerlik sınırlı içindeki bağımsız olunan herhangi bir parametreye göre grubun özelliğini belirlemesi için oluşturulmuştur (Raykov, Dimitrov, Li, Marcoulides ve Menold 2018). DMF ile ilgili tartışmaların bir sonucu olarak, yukarıda verilen Eşitlik 2 ve 3, iki parametreli MTK modelinin tahminlenenemeyen ikili maddeler grubu için doğru olması durumunda DMF için yeterli şartı
sağlamadığı göstermişti (Lord, 1980). DMF, bir test maddesine aynı yetenek düzeyinde olup farklı gruplardan gelen bireylerin maddeye doğru cevap verme olasılığının aynı olması durumunu ifade eder (Adams ve Rowe, 1988). DMF analizleri temelde, test puanlarının farklı gruplardan gelen değişkenliklerden etkilenip etkilenmediğini ve bu değişkenliklerin herhangi bir alt grup için bir fayda sağlayıp sağlamadığını araştırmayı amaçlar (Algina ve Crocke, 1986). Test ile ölçülen özellik farklı alt gruplara aynı ise, maddelerin aynı değişkenlik durumundan etkilendiğini ve aynı yetenek düzeyindeki bireylerin ölçülen ölçek yapısı da benzer yetenek düzeyinde olduğu söyleyebilir (Algina ve Crocke, 1986). Bu çalışmada kullanılan ikili Maddelerde ölçme değişmezliği analizi yöntemi, iki parametreli MTK altında DMF gösteren maddeleri belirleme yolu ile ölçme değişmezliğini test etme imkanı sunmuştur.

Araştırmanın Amacı

Bu çalışmanın genel amacı, PISA 2015 okuma becerileri alt testinin ana dili İngilizce olan ülkeler ile ana dili İngilizce olmayan ülkeler için dil becerileri açısından eşdeğerliğinin sağlanıp sağlanmadığının incelemesidir. Karşılaştırma ve değerlendirilmenin anlamalı olarak çalışılması, kültürler ve diller arasına ölçme eşdeğerliğinin olması, yani ana dili İngilizce olan ülkeler için herhangi bir yanlılık olmaması gerekmektedir. Belli bir kültürde ve dilde geliştirilmiş ölçükler, o kültür ve dilde özgü nitelik ve kavramsal özellikleri yansıtır. Bir ölçme araçının bir dilden başka bir dilde çevrilmesi, bu ölçüğün eşdeğer olduğunu garantisini vermez (Sireci ve Berberoğlu, 2000). Başka bir dille çevrilenecek ya da uyarlanacağ ise ölçme araçlarının orijinal formlarından farklı olacağını bilinmelidir. Söz konusu farklılıkların psikometrik, dil ve anlamlılık açılarından kabul edilebilir düzeyde olmaları sağlanmalıdır (Hambleton ve De Jong, 2003). Ulkelerin eğitim politikalarında büyük rol oynayan böyle bir araştırmada testlerin dil becerileri açısından kültürler arasıarı esdeğerliğinin sağlanması olması, doğru kararlar alınmasını ön koşullardan biridir (Arrffman, 2010; Baykal ve Cirici, 2010; He, Barrera-Pedemonte ve Bucholz, 2018). Tüm bu gereklikler için yapılacağı ölçümlerin titizlikle yapılması, ölçme araçının geçirilmiş olması için son derece önemlidir. Bu nedenlerle bu çalışmada PISA 2015 uygulamasının okuma becerileri testinin çeviri dil ve orijinal dil arasında herhangi bir ölçme değişmezliği sorunu olup olmadığını istatistiksel analiz yöntemleriyle incelenmiştir.

YÖNTEM

PISA 2015 uygulamasına İngiltere’den 14157, ABD’den 5712, Kanada’dan 20058, Japonya’dan 6647, Tayland’dan 8249 ve Türkiye’den 5895 öğrenci katılmıştır. Ülkelerin seçimi yapılırken ana dilleri ve okuma becerileri açısından başarı seviyelerini göz önünde bulundurulmuştur. Araştırma aynı dilde edilen altı ülkeden üç testi orijinal formda olan diğer üç ise ana dili İngilizce olan ülkeler için herhangi bir yanlılık olmaması dahil edilmiştir. Arastırma, dillerarasına etkilenen özellikli ölçüm olarak alınmış, doyu kararlar alınmasını ön koşullardan biridir (Arrffman, 2010; Baykal ve Cirici, 2010; He, Barrera-Pedemonte ve Bucholz, 2018). Tüm bu gereklikler için yapılacağı ölçümlerin titizlikle yapılması, ölçme araçının geçirilmiş olması için son derece önemlidir. Bu nedenlerle bu çalışmada PISA 2015 uygulamasının okuma becerileri testinin çeviri dil ve orijinal dil arasında herhangi bir ölçme değişmezliği sorunu olup olmadığını istatistiksel analiz yöntemleriyle incelenmiştir.
Tablo 1. Ülkeler Bazında Örneklem Frekansları

| Ülke  | N   | %    |
|-------|-----|------|
| Kanada| 524 | 34,4 |
| İngiltere| 384 | 25,2 |
| Tayland| 176 | 11,5 |
| Japonya| 145 | 9,5  |
| Türkiye| 159 | 10,4 |
| ABD   | 136 | 8,9  |
| Toplam| 1524| 100  |

Katılım oranlarına bakıldığında araştırma örnekleminin en fazla katılımcıya sahip ülke %34,4 ile Kanada ve en az katılımcıya sahip ülke ise %8,9 orani ile ABD olduğu anlaşılmaktadır. Araştırma örneklemini PISA 2015 uygulamasında katılan ülkelerden seçilmiş alt ülkeden 41 numaralı formu alan toplam 1524 öğrenci oluşturmaktadır. PISA 2015 okuma becerileri alt testi verilerine OECD’nin resmi web sayfasından ulaşılmıştır. Araştırma kapsamına alınan ülkeler PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkeler arasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamına alınan ülkelerden PISA 2015 uygulamasına bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katılan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kapsamında bilgisayar tabanlı değerlendirme ile katilan ülkelerarasından seçilmiştir. Araştırma kako...
serbest modelinde ve her bir madde bazında ayrı ayrı Bayesian Information Criterion (BIC) değerleri hesaplanmıştır. Parametre sayıları BIC üzerinde, bir diğer uyum indeksi olan Akaike Information Criterion (AIC) değerine göre daha büyük bir etkiye sahiptir (Frank J., Fabozzi ve Wiley, 2014). Nylund, Asparouhov ve Muthén (2007) bulguları doğrultusunda BIC’nin model iyiliği kriteri olarak kullanılmasına karar verilmiştir. Elde edilen her bir modelin BIC değeri ile M₀ serbest modelinin BIC değeri arasındaki farklar bulunmuştur. BIC değerleri arasındaki farkın (ΔBIC) büyüküğüne göre modelin uyumlu olduğu sonucu veriliyor. 6<ΔBIC<10 arasındaki değerler ise çok güçlü bir kanıt sunar. BIC değeri ne kadar küçüke model o kadar iyi demektir. M₀ serbest modeline göre daha büyük BIC farkı veren maddelerin modelle uyumu, bu farkın büyüküğüyle ters orantılı olarak azalmaktadır (Frank J., Fabozzi ve Wiley, 2014). Bu doğrultuda ΔBIC>10 olan maddeler oldukça kötü bir model uyumu veriyor anlamına gelir, dolayısıyla ölçme değişmezliği sağlanmadığına dair bir kanıt oluşturmaktadır.

**BULGULAR**

Araştırmada öncelikle yapı geçerliğine ilişkin bir kanıt elde etmek amacıyla her bir grup için ikili değişkenlerin yapısına uygun olarak DFA tamamlanmış, standartize edilmemiş faktör yükleri belirlenmiş ve model uyumu incelenmiştir. DFA ile elde edilen bulgular Tablo 2’de verilmiştir.

Tablo 2. PISA Okuma Becerileri Testine İlişkin Doğrulayıcı Faktör Analizi Sonuçları

| Grup (Ülkeler) | χ² | df | p    | n   | RMSEA | CFI | TLI |
|---------------|----|----|------|-----|-------|-----|-----|
| Anadili İngilizce Olan | 409.58 | 226 | .00  | 1044 | .03   | .96 | .97 |
| Anadili İngilizce Olmayan | 243.86 | 174 | .00  | 480  | .03   | .97 | .98 |

Tablo 2’deki model uyum indeksleri incelendiğinde anadili İngilizce olan grupta χ² değerinın (409.58), serbestlik derecesi (226) ile birlikte anlamlı olduğu görülmektedir (p < .05). Anadili İngilizce olmayan grupta da χ² değerinın (243.86), serbestlik derecesiyle (174) birlikte her iki grupta anlamlı olduğu tablodan anlaşılıyor (p < .05). RMSEA değerlerine bakıldığında, her iki grup için de .03 değerinde olduğundan modelin her iki grupta mükemmel uyum gösterdiğini söyleyebilir. CFI ve TLI uyum indeksleri incelendiğinde ise, ana dili İngilizce olan grup için CFI değerleri .96 ve TLI değerleri ise .97 ile kabul edilebilir uyum sağlamaktadır. Ana dili İngilizce olmayan gruptan CFI değerinin .97 ile kabul edilebilir ve TLI değer .98 ile iyi model uyumu verdiğini göstermektedir. DFA sonuçları PISA 2015 okuma becerileri testinden oluşan yapına her iki grup için de ayrı ayrı model uyumu gösterdiğini ortaya koymaktadır. Her bir gruba ait madde faktör yükleri ve a, b parametrelerinin farklılık göstermediğini incelemek için yapılan DFA analizi sonucunda elde edilen madde faktör yükleri, eşik değerler, a ve b parametreleri Tablo 3’te verilmiştir.
Tablo 3. PISA 2015 Okuma Becerileri Testi Dil Değişkenine Göre Oluşturulan Grupların DFA Sonuçlarına İlişkin Madde Parametreleri

| Madde | Anadili İngilizce Olan Ülkeler | Anadili İngilizce Olmayan Ülkeler |
|-------|-------------------------------|----------------------------------|
|       | \( \lambda \) | \( t \) | \( a \) | \( b \) | \( \lambda \) | \( t \) | \( a \) | \( b \) |
| 1     | 1.00 | -0.81 | .64 | -1.50 | 1.00 | -0.38 | .95 | -0.56 |
| 2     | 1.01 | -1.13 | .65 | -2.07 | .99 | -0.35 | .93 | -0.51 |
| 3     | 1.06 | -1.26 | .70 | -2.20 | .88 | -0.64 | .76 | -1.06 |
| 4     | 1.10 | -1.07 | .74 | -1.80 | .65 | -0.62 | .51 | -1.37 |
| 5     | 1.23 | -0.90 | .88 | -1.36 | .61 | -0.41 | .46 | -0.97 |
| 6     | 1.25 | -0.92 | .91 | -1.36 | 1.03 | -0.51 | 1.02 | -0.71 |
| 7     | 1.18 | .67 | .82 | 1.06 | 1.03 | .76 | 1.01 | 1.06 |
| 8     | 1.00 | .33 | .64 | .61 | .83 | .69 | .71 | -1.24 |
| 9     | 1.31 | -1.15 | .99 | -1.64 | .84 | -0.72 | .71 | -1.24 |
| 10    | .88 | .79 | .54 | 1.68 | .98 | .98 | .93 | 1.45 |
| 11    | 1.17 | -0.06 | .82 | -0.09 | .83 | .22 | .70 | .39 |
| 12    | .99 | -0.08 | .63 | -0.14 | .51 | -0.13 | .38 | -0.36 |
| 13    | 1.19 | -0.79 | .84 | -1.23 | .97 | -0.54 | .90 | -0.81 |
| 14    | .90 | -0.06 | .56 | -0.12 | .70 | .08 | .55 | .17 |
| 15    | .50 | .37 | .28 | 1.36 | .45 | .53 | .32 | 1.73 |
| 16    | .63 | -0.24 | .36 | -0.70 | .60 | .08 | .46 | .20 |
| 17    | 1.07 | -0.86 | .71 | -1.48 | .76 | -0.76 | .62 | -1.26 |
| 18    | 1.31 | -0.88 | .99 | -1.25 | .76 | -0.69 | .61 | -1.33 |
| 19    | .91 | -0.07 | .56 | -0.14 | .67 | .10 | .53 | .22 |
| 20    | 1.22 | -0.41 | .87 | -0.62 | 1.02 | .17 | .99 | .24 |
| 21    | .20 | -0.14 | .85 | -0.22 | 1.06 | .30 | 1.07 | .41 |
| 22    | .99 | -0.26 | .63 | -0.49 | .85 | -0.47 | .72 | -0.81 |
| 23    | .87 | -0.83 | .53 | -1.77 | 1.02 | -0.34 | .98 | -0.48 |
| 24    | .81 | .16 | .48 | .36 | .84 | .39 | .71 | .68 |
| 25    | .79 | -0.10 | .47 | -0.21 | .73 | .53 | .58 | 1.06 |
| 26    | .77 | -0.94 | .46 | -2.26 | .60 | -1.15 | .45 | -2.78 |
| 27    | 1.02 | -0.32 | .66 | -0.57 | .53 | -0.27 | .40 | -0.64 |
| 28    | 1.26 | .58 | .92 | .86 | .96 | .73 | .88 | 1.10 |

Not: \( \lambda \)= madde faktör yükü, \( t \)=madde eşik değeri, \( a \)=madde güçlük, \( b \)=madde zorluk
Bağdu Söyler, P., Aydın, B., Atılgan, H. / PISA 2015 Dil Gruplarına Göre Madde Parametreleri: İkili Değişkenlerle Ölçme Değişmezliği Çalışması

maddeler olduğu analiz edilmiştir. 15. madde ana dili İngilizce olmayan grupta en düşük faktör yüküne (.45) sahip maddedir ve diğer grubun 15. maddenin verdiği faktör yükü (.50) ile oldukça yakın değerdir. 12. maddenin faktör yükleri ve a parametreleri karşılaştırıldığında ana dili İngilizce olan grubun maddenin faktör yükü .99 ve a parametresi .63 iken ana dili İngilizce olmayan grupta maddenin faktör yükü .51 ve a parametresinin .38 olduğu görülmüştür. Bu değerler aynı özelliği ölçmesi gereken maddeler için birbirinden oldukça farklıdır. Her iki grubun faktör yükleri maddenin ana dili İngilizce olmayan grupta 4, 9, 18, 21, 27 ve 28. maddelerin faktör yüklerinin ve a parametrelerinin ana dili İngilizce olmayan grupta farklı değerler aldığını göstermektedir. Buradan bazı maddelerin her iki grup için de aynı şekilde çalışmadığı söylenebilir.

Madde eşik değerleri ve b parametreleri incelendiğinde 2. maddenin ana dili İngilizce olan grupta eşik değerinin -1.13 iken ana dili İngilizce olmayan grupta -.35 olduğu, b parametrelerinin ise ana dili İngilizce olan grupta -.07; ana dili İngilizce olmayan grupta ise -.51 olduğu görülmüştür. Bu değerler her iki grupta da aynı özelliği ölçmesi gereken bir madde için birbirinden oldukça farklı değerlerdir. Benzer şekilde 23. maddenin her iki grupta b parametrelerinin karşılaştırıldığında ana dili İngilizce olan grupta -.17, diğer grupta -.48 değerini aldığı görülmüştür. Bu doğrultuda, tablo 3’te 6, 8, 22, 25 ve 26. maddelere ait eşik değerleri ve b parametrelerinde farklılıklar olduğu görülmuştur. Toplamba 28 maddeden oluşan veri seti için 56 farklı modelde madde parametreleri ve BIC değerleri değişimi incelendirmiştir.

Ana dili İngilizce olan ve ana dili İngilizce olmayan her iki gruba ait iki değişkenli puanlanan madde maddeden çalışma değişmezliği analizi sonuçları Tablo 4’te verilmiştir. Tablo 4 ve Tablo 5 incelendiğinde her bir madde için madde faktör yükleri ve eşik değerlerinin serbest bırakılması ile 56 farklı modelden elde edilen BIC değerleri, bunların M₀ serbest modelindeki BIC değerlerin farkları (ΔBIC) ve madde faktör yükleri ile eşik değerleri yer almaktadır. M₀ serbest modelinin BIC değerleri ile her bir modelin BIC değerleri aynı ayrı karşılaştırılmıştır.
### Tablo 4. PISA 2015 Okuma Becerileri Testi Ölçme Değişmezliği Analizi Eşik Değerler

| Model | Par | BIC     | ΔBIC | ʎ  | t  | ʎ  | t  |
|-------|-----|---------|------|-----|----|-----|----|
| M0    | t0  | 44745.34 | -    | -   | -  | -   | -  |
| M1    | t1  | 44748.62 | 3.28 | -   | -1.69 | -1.27 |
| M2    | t2  | 44708.93 | -36.41* | - | -2.35 | -1.25 |
| M3    | t3  | 44737.70 | 7.64 | -   | -2.69 | -2.02 |
| M4    | t4  | 44748.02 | 2.68 | -   | -2.21 | -1.87 |
| M5    | t5  | 44746.98 | 1.64 | -   | -1.84 | -1.48 |
| M6    | t6  | 44752.57 | 7.23 | -   | -2.14 | -2.19 |
| M7    | t7  | 44739.41 | -5.93 | - | 1.66  | 1.00 |
| M8    | t8  | 44751.38 | 6.04 | -   | .71   | .87 |
| M9    | t9  | 44752.39 | 7.05 | -   | -2.76 | -2.66 |
| M10   | t10 | 44752.62 | 7.28 | -   | 1.65  | 1.61 |
| M11   | t11 | 44751.68 | 6.34 | -   | -0.10 | -0.24 |
| M12   | t12 | 44731.95 | -13.39* | - | -0.14 | -0.72 |
| M13   | t13 | 44749.29 | 3.95 | -   | -1.78 | -2.10 |
| M14   | t14 | 44748.84 | 3.50 | -   | -0.10 | -0.34 |
| M15   | t15 | 44752.65 | 7.31 | -   | .64   | .65 |
| M16   | t16 | 44748.44 | 3.10 | -   | -0.42 | -0.17 |
| M17   | t17 | 44749.87 | 4.43 | -   | -1.79 | -2.05 |
| M18   | t18 | 44745.93 | .59  | -   | -2.01 | -2.44 |
| M19   | t19 | 44751.14 | 5.80 | -   | -0.13 | -0.29 |
| M20   | t20 | 44742.71 | -2.63 | - | -0.84 | -0.37 |
| M21   | t21 | 44751.50 | 6.16 | -   | -0.21 | -0.71 |
| M22   | t22 | 44691.77 | -53.57* | - | 1.13  | -1.58 |
| M23   | t23 | 44745.36 | .02  | -   | -1.66 | -1.30 |
| M24   | t24 | 44752.56 | 7.22 | -   | .31   | .27 |
| M25   | t25 | 44722.68 | -22.66* | - | -1.15 | .58 |
| M26   | t26 | 44725.27 | -20.07* | - | -1.75 | -2.66 |
| M27   | t27 | 44743.97 | -1.37 | -   | -0.59 | -0.97 |
| M28   | t28 | 44743.49 | -1.85 | -   | 1.32  | .82 |

Not: ʎ= madde faktör yükü, t=madde eşik değeri
Grup 1: Ana dili İngilizce Olan Ülkeler   Grup 2: Ana dili İngilizce Olmayan Ülkeler
Tablo 5. PISA 2015 Okuma Becerileri Testi Ölçume Değişmezliği Madde Faktör Yükleri

| Model | Par | BIC      | ΔBIC  | $\hat{\lambda}$ | t  | $\hat{\lambda}$ | t  |
|-------|-----|----------|-------|-----------------|----|-----------------|----|
| M29   | $\lambda_1$ | 44747.36 | 2.02  | 1.15            | -  | 1.57            | -  |
| M30   | $\lambda_2$ | 44722.44 | -22.90* | 1.09          | -  | 2.09            | -  |
| M31   | $\lambda_3$ | 44741.84 | -3.50  | 1.20           | -  | 1.73            | -  |
| M32   | $\lambda_4$ | 44752.67 | 7.43   | 1.19           | -  | 1.20            | -  |
| M33   | $\lambda_5$ | 44751.06 | 5.72   | 1.36           | -  | 1.17            | -  |
| M34   | $\lambda_6$ | 44752.55 | 7.21   | 1.68           | -  | 1.74            | -  |
| M35   | $\lambda_7$ | 44749.53 | 4.19   | 1.30           | -  | 1.74            | -  |
| M36   | $\lambda_8$ | 44752.60 | 7.26   | 1.09           | -  | 1.14            | -  |
| M37   | $\lambda_9$ | 44752.26 | 6.92   | 1.75           | -  | 1.65            | -  |
| M38   | $\lambda_{10}$ | 44746.56 | 1.12   | .92            | -  | 1.49            | -  |
| M39   | $\lambda_{11}$ | 44748.68 | 4.34   | 1.40           | -  | 1.05            | -  |
| M40   | $\lambda_{12}$ | 44731.21 | -24.13* | 1.06       | -  | .44            | -  |
| M41   | $\lambda_{13}$ | 44751.05 | 5.71   | 1.52           | -  | 1.33            | -  |
| M42   | $\lambda_{14}$ | 44752.08 | 6.74   | .89            | -  | .78            | -  |
| M43   | $\lambda_{15}$ | 44752.64 | 7.30   | .50            | -  | .51            | -  |
| M44   | $\lambda_{16}$ | 44750.76 | 5.42   | .59            | -  | .77            | -  |
| M45   | $\lambda_{17}$ | 44749.77 | 4.43   | 1.28           | -  | 1.05            | -  |
| M46   | $\lambda_{18}$ | 44736.14 | -9.20* | 1.75       | -  | 1.15            | -  |
| M47   | $\lambda_{19}$ | 44751.27 | 5.93   | .94            | -  | .77            | -  |
| M48   | $\lambda_{20}$ | 44749.49 | 4.15   | 1.46           | -  | .86            | -  |
| M49   | $\lambda_{21}$ | 44751.57 | 6.23   | 1.46           | -  | 1.69            | -  |
| M50   | $\lambda_{22}$ | 44736.79 | -8.55* | 1.18       | -  | .63            | -  |
| M51   | $\lambda_{23}$ | 44734.46 | -10.88* | .97       | -  | 1.62            | -  |
| M52   | $\lambda_{24}$ | 44748.02 | 2.68   | .75            | -  | 1.11            | -  |
| M53   | $\lambda_{25}$ | 44747.32 | 1.98   | .78            | -  | 1.17            | -  |
| M54   | $\lambda_{26}$ | 44739.60 | -5.71  | .95            | -  | .44            | -  |
| M55   | $\lambda_{27}$ | 44736.40 | -8.64* | 1.12       | -  | .58            | -  |
| M56   | $\lambda_{28}$ | 44752.64 | 7.30   | 1.43           | -  | 1.40            | -  |

Not: $\hat{\lambda}$= madde faktör yükü, t=madde eşik değeri, Grup 1: Ana dili İngilizce Olan Ülkeler Grup 2: Ana dili İngilizce Olmayan Ülkeler

$M_0$ serbest modelinde BIC değeri 44745.34 olarak bulunmuştur. Her bir modeldeki BIC değerinin bu değerle farkı hesaplanmıştır. Elde edilen bulgulara göre, 2. maddenin Model 2 de $\Delta$BIC değerinin -
36.41 (ΔBIC>10) ve Model 30’dan ΔBIC değeri -22.90 (ΔBIC>10) olduğu görülmüştür. 2. maddeye ait eşik değerlerinin de ana dili İngilizce olan ülkelerden oluşan grup (Grup 1) için -2.35 ve ana dili İngilizce olmayan ülkelerden oluşan grup için (Grup 2) -1.25 değerleriyle birbirinden oldukça farklı olduğu analiz edilmiştir. Bu doğrultuda 2.addenin model uyumu göstermediği ve her iki grup için de karşılaştırılabilir olmadığı söylenebilir. DüDV 18’in Model 46’dağlı ΔBIC değerenin -9.20 (6<ΔBIC<10) ile yine kötü model uyumu veren aralıktaki olduğu analysiz edilmiştir. Madde eşik değerleri ve a, b parametrelerinin de birbirinden farklı değerleri aldığını Tablo 3’ten gorelimaktedir. Benzer şekilde madde 22’nin Model 22’deki ΔBIC değerinin -53.57 (ΔBIC>10) ve Model 50’de bu değerin -8.55 (6<ΔBIC<10) olduğu görlümüştür. Bu maddelere ait parametrelerin de her iki grup bazıdan birbirinden farklı olduğu yine Tablo 3’ten anlaşılmaktadır. Benzer şekilde 12, 23, 25, 26 ve 27. maddelerin de kötü uyum verdiği görülmüştür.

Elde edilen bulgulara göre 28 maddeden oluşan PISA-2015 okuma becerileri testinin 8 maddesinin ölçme değişmezliğini tam olarak sağlamadığı söylenebilir. İlkli puanlanmış maddelerde ölçme değişmezliği analizi ile ulaştılan sonuçlarda en az bir maddeinin ölçme değişimliğini tam olarak sağlamaması, testin ölçme değişimliğini sınırladığı testin bütününde bir ölçme değişimliğini sorun olabileceğini ortaya koymaktadır (Raykov ve ark., 2018).

**SONUÇLAR ve TARTIŞMA**

Bu araştırımada PISA 2015 okuma becerileri testinin ana dili İngilizce olan ülkeler ile ana dili İngilizce olmayan ülkeler arasında dil değişikleri açısından ölçme değişimliğini ikili puanlanmış maddelerin yapsına uygun olarak test edilmişdir. Veri setinin ikili kategorik yapsına uygun olarak ölçme değişimliğini analiz edilmişdir. Ana dili İngilizce olan ve ana dili İngilizce olmayan ülkelerden oluşan iki grup için ayrı ayrı DFA yapılara model uyumu incelmemiş, faktör yapılacakların her bir grup için doğrulandığı sonucuna ulaşılmıştır. DFA ile elde edilen bulgularla her bir maddeinin ana dili İngilizce olan ve ana dili İngilizce olmayan gruplarında faktör yükleri ve eşik parametreleri karşılaştırılmıştır. PISA 2015 okuma becerileri testinin her iki grupta aynı özelliği olan varyant maddelerin bazarlarının faktör yüklere ve eşik parametrelerine birbirinden oldukça farklı olduğu analiz edilmiştir. Dolayısıyla bu durumun, ana dili değişkenine göre oluşturulan grupların birbirleriyle karşılaştırılabilirliği için bir sınırlık olabileceğini sonuçuna varılmıştır.

Gruplar arasında ölçme değişimliğini incelemek için ikili puanlanmış maddelerin yapsına uygun analizler yapılmıştır. Elde edilen bulgulara göre DFA sonuçlarına paralel olarak, ölçme değişimliğini sağlayan dününilen maddelerde serbest modele (M0) göre BIC değerlereinin daha büyük olduğu, bu farklıların 6 ile 10 arasında da 10’dan büyük olduğu dolayısıyla modele uyumunun azaldığı belirlenmiştir. Bu maddele ilgili madde eşik değerleri ve faktör yükleri karşılaştırılmıştır. Model uyumu düşük olan maddelerin birçoğunda madde eşik değerleri ve faktör yüklerinin de birbirinden farklı olduğu analiz edilmiştir. Bu doğrultuda PISA 2015 okuma becerileri testinin 41 numaralı formunda yer alan 28 maddeden 8’inin ana dili değişkenine göre oluşturulan gruplar arasında karşılaştırma yapılması ciddi şekilde sınırlığı sonucuna varılmıştır. En az bir maddeye böyle bir sınırlık olması testin bütününü için ölçme değişimliğini sınırladığı tam olarak sağlanamadığı anlamına gelmektedir (Raykov ve ark., 2018). Dolayısıyla bu testte ölçme değişimliğini, ana dili İngilizce olan ülkeler ile ana dili İngilizce olmayan ülkelerden oluşan gruplar arasında karşılaştırılabilir olmasına kısıtlamalar getiren kaynaklar tanımlanmadan tam olarak yorumlanamayacağı yarışmasına varılmıştır. Alanın üzerinde çalışan bulguların ulaştığı sonuçlarla aynı doğrultuda olan ölçme değişimliğini çalışmalarını bulunmaktadır, bir diğer ifade ile, çalışma bulguları alanyazında tutarlıdır. Örneğin, Ercikan ve Koh (2005), 1995 yılına uygulan TIMSS verilerini kullanarak İngilizce ve Fransızca versiyonlarının ölçme değişmezliğini araştırmışlardır. Araştırma Kanada’dan kadının öğrencileri dahil edilmiştir. Madde tepki kurum ve çok gruplu dolayı faktör analizi ile karşılaştırılar yapmış, matematik ve fen alanındaki genel performansların karşılaştırıarak için kullanmada ciddi sınırlamaların olduğunu göstermiştir. Öğretmen (2006), ülkemizde de katkıda PIRLS 2001

ISSN: 1309 – 6575  Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
kapsamında uygulanan okuma parçaları testlerinin psikometrik özelliklerini ABD ve Türkiye örneklemelerinde karşılaştırılarak olarak incelemiştir. Araştırma iki aşamada gerçekleştirilmiştir olup, birinci aşamada okuma parçaları testlerinin öçtüğü düşünen yapıların kültürlerde göre eşdeğer olup olmadığını ÇGDFA yöntemi ile test edilmiştir. İkinci aşamada ise test maddelerinin kültürlerde göre DMF içeriği bir şekilde, bazı maddelerin karşılaştırılması ve olabildirlik oran testi ile yapılan analizler sonucunda maddelerinin yoğun olarak DMF içeriği olduğu belirtilmiştir. Asıl ve Gelbal (2012), PISA 2006 kapsamında uygulanan okuma parametrelerinin kültür ve dil bakımdan ölçme değişmezliğini sınırlayan maddelere ulaşılamadığı için araştırmaya dahil edilememiştir. Çalışmaya Avustralya, Yeni Zelanda, Amerika Birleşik Devletleri ve Türkiye örneklemeleri dahil edilmiştir. Çağrı çevremesi, kullanılan ölçme yöntemlerinde düşüncesel maddelerin kültür ve dille göre DMF gösteren göstermediği araştırılmıştır. ulaştılan sonuçlara göre, ülkeler arasında kültür ve dilsel açıdan farklılıklar arttıkça doğru orantılı olarak değişen madde fonksiyonu gösteren madde sayısının da arttığı gözlemlenmiştir.

PISA gibi bir çok dille çevrili yapılan uluslararası uygulamaların orijinal hali İngilizce dilinde yazılan sorular, uzman çevirmenler tarafından başka bir dille çevrilip ardından orijinal versiyonu ile eşdeğerliğini değerlendirmek için tekrar çevrilmektedir. Bu uygulamalarla yalnızca çeviriğin kaynaklı değil, esdeğeriğin sorunu olup olmadığını veya yanlışlık kaynağını özellikle etmenler, karşılıkların eşdeğerliği kavramı ve.childNodes olanlar belirlenmeli. Bu etmenleri dikkatlice incelemek için ölçme yönteminin, diller arasındaki farklılıkları kültür farklılıklarından kaynaklanan ve dil yansıyan etkileri hakkında bilgi edinilmedir (Goldstein, 2017). farklı diller ve kültürler arasında yapılan değerlendirmelerde bu değişkenlere bağlı yanlışlık gösteren ya da esdeğerliği sağlamayan madde içerikleri mutlaka göz önünde bulundurulmalıdır. Belli bir dille özgü olan ve ayrıcalık sağlayacak anlatımlar içeren maddeler testin dışında bırakılmalıdır.

Sorular erişime açık olsaydı, ölçme değerişme zorunluluğu getiren maddeler incelenerek sonuçlar arasındaki farklılıklar ayrıntılı olarak görülebilir. PISA’nın resmi internet sitesinde bu maddelere ulaşlamadığı için araştırılamaya da dahil edilememiştir. PISA uygulamalarında maddelerin erişime açık olması ve bir çok değişkenin açısından içerik olarak incelemesi, ölçme değişmezliğini sınırlayan faktörlere araştırılması açısından önemlidir. Bu doğrultuda PISA’nın resmi sitesinde yalnızca uygulandığı yila özgü alanları soruları değil, tüm soruların erişime açık olması önerelimidir.

KAYNAKÇA
Adams, R., & Rowe, K. (1988). Educational research, methodology, and measurement: An international handbook. Oxford: Pergamon Press.
Akbaş, U. ve Tavşancıl, E. (2015). Farklı özellik büyüklüklerinde ve kayıp veri örüntülerinde ölçülere psikometrik özelliklerinin kayıp veri başına etkisi teknikleri ile incelemesi., Journal of Measurement and Evaluation in Education and Psychology, 6(1), 38-57.
Algina, J. & Crocker, L., (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.
Arffman, I. (2002). In search of equivalence: Translation problems in international literacy studies. Finland.
Arffman, I. (2010). Equivalence of translations in international reading literacy studies. Scandinavian Journal of Educational Research, 54(1), 37-59.
Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. Education and Science, 236-249.
Asil M., & Brown, G. (2015). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. International Journal of Testing, 16(1), 71-93.
Baker, F. B. (2016). The basics of item response theory. Ankara: Pegem Academy.
Baykal, A., & Cirici, R. (2010). Item revision to improve construct validity: A study on released science items in Turkish PISA 2006. Procedia Social and Behavioral Sciences, 2(2), 1931-1935.
Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.
Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. Assessment in Education: Principles, Policy & Practice, 9(3), 387-399.
Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: Guildford.
Cheema, J. (2012). Handling missing data in educational research using SPSS. Yayınlanmamış Doktora Tezi, George Mason University.
Downey, R., & King, C. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, 175-191.

Elosua, P. (2011). Assessing Measurement Equivalence in Ordered-Categorical Data. *Psicológica*, 403-421.

Enders, C. K. (2010). *Applied missing data analysis*. (1. Ed.). New York: The Guilford Publications, Inc

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMMS. *International Journal Of Testing*, 23-3

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting test for use in other languages and cultures. *APA Handbook of Testing and Assessment in Psychology* (s. 545-569). içinde Washington: American Psychological Association.

Frank J. Fabozzi, S. M., & Wiley , J. (2014). Model Selection Criterion: AIC and BIC. *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*.

French, B. F., & Finch, W. H. (2006). Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance. *Structural Equation Modeling*, 13(3), 378-402.

Goldstein, H. (2017). Measurement and evaluation issues with PISA. *Routledge*.

Gregoric, S. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 78-94.

Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3), 249-266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *ERI Monograph Series: Issues and Methodologies In Large-Scale Assessments*, 2, 63-84.

Hambelton, R. K., & Swamimathan , H. (1985). *Item Response Theory*. Nijhoff Publishing.

Hambelton, R. K., & De Jong, J. A. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing*, 127-134.

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

Herdman M., Rushby J. F., & Badia X. (1998). *A Model of Equivalence in The Cultural Adaptation of HRQol Instruments: The Universalist Approach*. Quality Of Life Research, 7(4), 323-335.

He, J., Barrera-Pedemonte, F., & Bucholz, J. (2018). Cross-cultural comparability of non-cognitive constructs in TIMSS and PISA. *Assessment in Education:Principles, Policy & Practice*, 26(4), 369-385.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*.

Jöreskog, K.G., Sörbom, D., Du Toit, S.H.C., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd ed.). Lincolnwood, IL: Scientific Software International.

Kankaras, M., & Moors, G. (2013). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, 43(3), 381-399.

Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling*, 212-228.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford publications.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 210-231.

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

Lubke, G. H., & Muthén, B. O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons . *Structural Equation Modeling A Multidisciplinary Journal*, 11(4), 514-534.

Martin, M., Mullis, I., Gonzalez, E., Gregory, K., Smith, T., Chrostowski, S., O’Connor, K. (2000). TIMSS 2009 International Science Report: Findings from IEA’s Repeat of The Third International Mathematics and Science Study at the Eighth Grade.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361-388.

Milli Eğitim Bakanlığı (2016). *PISA 2015 International report*. Ankara. Erişim adresi: https://odsgm.meb.gov.tr/www/2015-pisa-uluslaraporu/icerik/204

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*.

Millsap , R. E. (2011). *Statistical approaches to measurement invariance*. New York: US: Routledge/Taylor & Francis Group.
Bağdu Söyler, P., Aydin, B., Atılgan, H. / PISA 2015 Dil Gruplarına Göre Madde Parametreleri: İkili Değişkenlerle Ölçme Değişmezliği Çalışması

Muthén, B., and Asparouhov, T. (2013). BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17. Available online at: www.statmodel.com

Muthén , B., Asparouhov, T., & Morin, A. J. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. Journal Of Management.

Muthén, L. K., & Muthén, B. O. (2016). Mplus user’s guide. Los Angeles, CA: Muthén & Muthén.

Nylund, K.L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. Structural Equation Modeling, 14, 535-569.

Organisation for Economic Co-operation and Development (2016). Erişim adresi: http://www.oecd.org/education/Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. Journal of Research in Personality, 40(4), 411-423.

Öğretmen, T. (2006). Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği. Ankara.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. Journal of Psychological Test and Assessment Modeling, 53(3),315-333.

Onen, E. (2009). Ölçme değişmezliğinin yapısal eşitlik modellemesi teknipleri ile incelenmesi. Doktora Tezi, Ankara Üniversitesi, Ankara.

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. Educational and Psychological Measurement, 73(4) 713-727.

Raykov, T., Dimitrov, D., Marcoulides, G., Li, T., & Menold, N. (2018). Examining Measurement Invariance and Differential Item Functioning With Discrete Latent Construct Indicators: A Note on a Multiple Testing Procedure. Educational and Psychological Measurement, 78(2), 343-352.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552-566.

Rubin, D. B., (1976). Inference and missing data. Biometrika, 63, 581-592.Salzberg, T., Sinkovics, R., & Schlgelmich, B. (1999). Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches. Australasian Marketing Journal, 23-38.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. Applied Measurement in Education, 13(3), 229-248.

Sirganci, G., Uyumaz, G., & Yandi, A. (2020). Measurement invariance testing with alignment method: Many groups comparison. International Journal of Assessment Tools in Education, 7(4), 657-673.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52(3), 393-408.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. Frontiers in Psychology, 4, 1–15.

Wu, D., Li, Z., & Zumbo, B. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. Practical Assessment, Research & Evaluation, 12(3), 1-26.
Ek A. Anadili İngilizce Olan Ülkelerde DFA İçin Kullanılan Mplus 8.0 Kodları

TITL̦E: this is an example of a CFA with categorical factor indicators
DATA: FILE IS ING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;
Ek B. Anadili İngilizce Olmayan Ülkelerde DFA İçin Kullanılan Mplus 8.0 Kodları

TITLE: this is an example of a CFA with categorical factor indicators
DATA: FILE IS NONING.dat;
VARIABLE: NAMES ARE u1-u28;
CATEGORICAL ARE u1-u28;
MISSING ARE ALL(999);
MODEL: f1 BY u1-u28;
Ek C. İkili Puanlanmış Maddelerde Ölçme Değişmezliği İçin Kullanılan Mplus 8.0 kodları

M₀ Serbest Modeli İçin Yazılan Kod:

TITLE: Raykov (2018) M0
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1 @ 1;
%C#2%
 f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;

Madde Eşik Değerlerinin Serbest Birakıldığı Bir Model Örneği (M₁-M₂₈):

TITLE: Raykov (2018) M1 (relase first threshold)
!LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
 f1 BY u1* (L1)
u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1 @ 1;
%C#2%
 f1 BY u1* (L1)
u2-u28 (L2-L28);
[u2$1-u28$1](T2-T28);
[u1$1*];
[f1*];
f1*;
Madde faktör Yüklerinin Serbest Bırakıldığı Bir Model Örneği(M29-M56):

TITLE: Raykov (2018) M29 (relase first loading)
!LISTWISE=ON;
DATA: FILE = multicfaALL1.dat;
VARIABLE: NAMES = g u1-u28;
CATEGORICAL = u1-u28;
KNOWNCLASS = C(g = 1 g = 2); !g=1 ING, g=2 NOing
CLASSES = C(2);
MISSING=ALL(999);
ANALYSIS: ESTIMATOR = ML;
TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
MODEL:
%OVERALL%
 f1 BY u1* (L1)
 u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1@0];
f1@1;
%C#2%
 f1 BY u1*
 u2-u28 (L2-L28);
[u1$1-u28$1](T1-T28);
[f1*];
f1*;