EJAES

JOURNAL OF ADVANCED EDUCATION STUDIES
İleri Eğitim Çalışmaları Dergisi
4(2): 88-105, 2022

MEASUREMENT INVARIANCE OF TURKISH "CENTRAL EXAM FOR SECONDARY EDUCATION" BY SPECIFIC LEARNING DISABILITY

Selma ŞENEL[1]*

Geliş Tarihi/Received:24.06.2022 Kabul Tarihi/Accepted:16.11.2022 Elektronik Yayın/Online Published:15.12.2022

DOI: 10.48166/ejaes.1135479

ABSTRACT

Ensuring measurement invariance for students with disabilities is critical for fair measurement in large-scale testing. Specific learning disability constitutes the largest group among disability groups. In this study, it was aimed to examine the measurement invariance of the Turkish Central Exam for Secondary Education according to whether or not students have a specific learning disability. 994 students diagnosed with specific learning disability formed the focus group, whilst 1,000 students without any disability constituted the reference group. Mantel Haenszel and Lord's chi-square methods were used to determine whether or not the items in each subtest showed Differential Item Functioning (DIF). In addition, by applying Multigroup Confirmatory Factor Analysis, the configural invariance, metric invariance, scalar invariance, and strict invariance of the subtests were examined. The study's findings proved that 34 of the 90-item test indicated DIF according to both methods. Eleven items show moderate DIF and five show high DIF. Metric invariance is not provided in all subtests, with factor loadings in all subtests varied between the groups.

Keywords: Differential item functioning; factorial invariance; measurement invariance; specific learning disability; test accommodations

[1]Department of Educational Sciences, Faculty of Education, Balıkesir University, Turkey e-mail: selmahocuk@gmail.com, selmasenel@balikesir.edu.tr, ORCID:0000-0002-5803-0793
*This study was presented as an oral presentation at the 7th International Congress on Measurement and Evalluation in Education and Psychology.

ORTAÖĞRETIME GEÇIŞ MERKEZI SINAVININ ÖZEL ÖĞRENME GÜÇLÜĞÜ OLAN ÖĞRENCILERE GÖRE ÖLÇME DEĞIŞMEZLIĞININ İNCELENMESI

ÖZET

Geniş ölçekli testlerin, özel gereksinimli öğrenciler için ölçme değişmezliğinin sağlanması ve buna yönelik bilimsel analizler adil ölçmeler için kritiktir. Özel öğrenme güçlüğü, özel gereksinim grupları içerisinde en büyük grubu oluşturmaktadır. Bu araştırmada ortaöğretime geçiş sınavının öğrencilerin özel öğrenme güçlüğü olup olmama durumuna göre ölçme değişmezliği incelenmiştir. Araştırmada öğrenme güçlüğü tanısı olan 994 öğrenci odak grubu, özel gereksinimi olmayan 1000 öğrenci ise referans grubu oluşturmuştur. Her bir alt testteki maddelerin Değişen Madde Fonksiyonu(DMF) gösterip göstermediği Mantel Haenszel ve Lord'un ki karesi yöntemleri ile incelenmiştir. Bunun yanında, Çoklu Grup Doğrulayıcı Faktör Analizi uygulanarak, alt testlerin yapısal, zayıf, güçlü ve katı değişmezlikleri aşamalı olarak incelenmiştir. Araştırma sonucuna göre 90 maddelik testin 34 maddesi her iki yönteme göre DMF göstermektedir. On bir madde orta ve beş madde ise yüksek düzeyde DMF göstermektedir. Zayıf değişmezlik ise tüm alt testlerde sağlanmamaktadır. Bu sonuca göre, sınavın tüm alt testlerinde, faktör yükleri gruplar arasında değişiklik göstermektedir. Bu sonuçlara göre söz konusu sınavın özel öğrenme güçlüğüne göre ölçme değişmezliğini sağlamadığı belirtilebilir.

Anahtar Kelimeler: Değişen madde fonksiyonu; faktöriyel değişmezlik; ölçme değişmezliği; özel öğrenme güçlüğü; test düzenlemeleri

1. INTRODUCTION

The results obtained from tests applied to individuals with disabilities and their peers should be comparable. This is essential for the validity of test scores and for the fairness of decisions that are subsequently taken based upon the results of the measurement. According to international measurement standards (AERA et al., 2014), an individual's disability, that is, a feature that is not intended to be measured, should not interfere with the measurement result of the intended construct. In addition, when tests are applied to groups with different characteristics, such as students with disabilities, the necessity of presenting and reporting evidence of the validity of the test in these separate groups should be reported. Especially in large-scale tests, in which scores are the basis of critical decisions made that affect the lives of the test takers, measurement invariance plays a critical role for disability groups.

If the aim of a test is not to measure an individual's disability or a construct related to disability, the disability itself should therefore not affect the test results. Getting assistance from a reader/coder (person or device), taking additional time where permitted, and presenting a larger font size are some of the test accommodations that are used to prevent test results from being affected by the disability of the test takers (Cortiella, 2005). There have been numerous research studies that have focused on the extent to which test accommodations ensure/increase the validity of the measurement (Bolt & Ysseldyke, 2008; Buzick & Stone, 2011; Elbaum, 2007; Gregg & Nelson, 2012; Lai & Berkeley, 2012; Lindstrom & Gregg, 2007; Middleton & Laitusis, 2007; Rogers et al., 2014, 2016; Stone et al., 2010). Today, test accommodations, whose current positive effects on validity have been largely proven by validated

research, are now widely applied in large-scale tests. However, the effective practice of test accommodations does not guarantee measurement invariance according to disability groups. Research has proven that tests conducted with accommodations do not, or rarely, provide measurement invariance for disability groups (Knickenberg et al., 2020; Şenel, 2021; Yılmaz, 2019).

Measurement invariance is the condition that the individual's group membership, which is unrelated to any characteristic being measured in the test, has no effect on the score or outcome (Mellenbergh, 1989). In this study, measurement invariance is where test items or test scores are not dependent on knowing if a test taker has any disability or whether they have received any form of test accommodations. In the absence of measurement invariance, tests treat groups differently, and therefore the validity of a test score that favors a certain group may be seen as skeptical. In this case, the scores from different groups in the same test cannot be readily compared (Borsboom, 2006).

Measurement invariance can be examined at the individual test item level or at the test level itself. Differential Item Functioning (DIF) and Differential Distractor Functioning (DDF) are forms of analysis that are frequently used in item level examinations (Abedi et al., 2007; Finch & French, 2007; Mori et al., 1974; Stone et al., 2010). What we usually understand from *measurement invariance* is invariance based on the test score. In other words, the measurement model should have the same construct across more than one group. In order to monitor and control test level *measurement invariance*, it is expected that the factor loadings of the test items, correlations between factors, and also error variances are the same between groups (Van De Schoot et al., 2015).

Specific learning disability (SLD) is a neurodevelopmental disorder that may negatively affect the individual's listening, speaking, reading, writing, spelling, concentration, mathematics, reasoning, motor, and organizational skills (Kavale & Forness, 2000; Kishore et al., 2021). SLD is characterized by unexpected low performance in certain academic fields, even though the absence of intellectual disability, sensory impairment, emotional disturbance, cultural deprivation, and insufficient instruction (Büttner & Hasselhorn, 2011). Students with SLD have the highest rate among all students with disabilities (National Center for Statistics Education, 2021). The literature reports that between 5% and 15% of the school-age population have some SLD (Bolt & Ysseldyke, 2008; Elliott et al., 2018; First, 2013; Grigorenko et al., 2019; Rogers et al., 2019). As a result, the rate of individuals with disabilities participating in large-scale testing is considered to be high, and a significant portion of measurement invariance studies are conducted with individuals who have some form of SLD, and varies depending on the disability or the use of test accommodations (Rogers et al., 2014, 2016, 2019).

Validity evidence concerns regarding the test results of students with SLD can be addressed in different dimensions (Bolt & Ysseldyke, 2008): (1) Do the items cause students to experience difficulties due to their learning disabilities? (2) Does the preferred accommodation affect the measurement of the structure? (3) Is the preferred accommodation considered sufficient? (4) Does reading fluency affect test scores? and (5) If it is a test consisting of open-ended items, can the writing difficulties of test takers affect the test score? Validity concerns may vary and increase with research.

In studies dealing with the validity of tests administered to students with SLD, the focus of research is the effect of test accommodations. Students with SLD generally make use of test accommodations related to the presentation of the tests (Rogers et al., 2014). Extended time is the most frequently used and is considered a significant accommodation that individuals with SLD opt to use (Camara et al., 2005; Gregg & Nelson, 2012; Kingsbury & Houser, 1988; Koretz, 1997). The read-aloud accommodation is frequently preferred; however, the literature has not provided a clear picture of the impact of the read-aloud accommodation, with varied results having been published to date. While some studies indicate significant increase in favor of individuals with SLD (Brumfield, 2014; Fletcher et al., 2006), others have reported increased scores for individuals without disabilities (Elbaum, 2007; Elbaum et al., 2004), or proving similar increases in the scores of those with and without a disability (Meloy et al., 2000). As understood from the current literature; although it is necessary and appropriate to provide test accommodations for students with physical or sensory disabilities, test accommodations and their effects are considered to be more controversial for students with SLD (Bolt, 2004). It is also observed that there are significant practice differences with regards to test accommodations between different countries, and even between different states in the example of the United States (Lai & Berkeley, 2012). However, the validity of tests for students with SLD, which has the highest rate among all need groups, is a situation that should always be examined and reported on the basis of accountability.

The literature that has focused on examining measurement invariance for SLD has mostly employed test-based factorial invariance analyses. Cook et al. (2010) compared the basic factors in the measurement of individuals with and without SLD in a Fourth Grade State-Standards-based English Language Arts (ELA) assessment, which consists of a total of 75 multiple-choice items, including both reading and writing parts. Results of the factorial analysis indicated that the test provided measurement invariance. Similarly, Steinberg et al. (2011) examined fifth-grade science test scores of students with and without SLD according to state standards. They examined factorial invariance with test-level exploratory and confirmatory factor analyses, along with item-level analysis. The findings confirmed the validity of the test scores. Another study that confirmed the provision of measurement invariance between students with and without SLD through examining factorial invariance was the work of Randall and Engelhard (2010), who employed both confirmatory factor analysis and the Rasch model in their study. In another study, Kim et al. (2009) also found that a statewide secondary school science test provided factorial invariance between groups of students with and without SLD.

Along with the studies that proved non-invariance between groups, there have also been studies that have examined item-based measurement invariance and which have shown that test results were in favor of a certain group. Kamata and Vaughn (2004) examined whether or not a 40-item statewide math test showed DIF for individuals with SLD. According to the result of the DIF analysis conducted using the Mantel-Haenszel and logistic regression methods, it was found that three items showed DIF against students with SLD and one item against students without SLD. Anjorin (2009) also researched DIF according to disability status in a statewide high-stake math test administered in the spring of 2003 to

students seeking a high school diploma in the United States. The study's results proved that items showing DIF worked in favor of those individuals without a disability.

In Turkey, studies examining the measurement invariance of large-scale tests according to the disability status of test-takers (Ozarkan et al., 2017; Şenel, 2021; Yılmaz, 2019) have also been quite limited in number. Şenel (2021) assessed measurement invariance of the Turkish "Central Exam for Secondary Education Institutions" (Milli Eğitim Bakanlığı[MEB][Turkish Ministry of National Education], 2018) in terms of visually impaired students. This large-scale test is considerable importance, since its results are used to decide on students' transition from secondary school to high school. The study's results highlighted that 17.78% of test items indicated DIF, and that 62.5% of the DIF detected items represented some form of disadvantage for students with visual impairments. Yılmaz (2019) also examined bias in terms of the mathematics (n = 20) and science (n = 20) subtests in 2016-2017 form of the test, according to disability status. Three disability groups; visually impaired, hearing impaired, and physically disabled students were compared with each other in pairs in Yılmaz's (2019) study. The results showed that four items in the mathematics subtest and eight items in the science subtest were found to be biased. In another study, Ozarkan et al. (2017) examined the DIF of items from the mathematics subtest (n = 20) in the 2015-2016 first semester exam according to the visual disability status of the participant individuals. Their findings showed that all 14 mathematics items, that provided analyses assumptions, contained negligible levels of DIF.

Studies on the validity of the results of the tests taken by individuals with SLD, whose rates in society are deemed quite high, have increased at a certain level over the past two decades. However, as the literature reports, the validity of these tests in which individuals with SLD and all other disability groups had participated was not adequately questioned during the development of the tests. The validity of large-scale test results, which are effective in critical decisions in the lives of individuals, is lower for these special groups and as a result may directly harm social justice. Although various test accommodations are provided, they alone do not guarantee measurement invariance. The first step in ensuring validity evidence of tests administered to students with disabilities is to conduct exploratory research on the tests and to report the research findings. Students with SLD, who have a high percentage of students with disabilities, should be given priority. Therefore, the aim of the current study is to examine the measurement invariance of the Central Exam for Secondary Education applied throughout Turkey according to students with SLD. The *Central Exam for Secondary Education* is a central examination administered annually since 2018 by the Turkish Ministry of National Education. The purpose of the exam is the placement of eighth-grade students to different high schools such as science high schools, social sciences high schools, and vocational and technical high schools (MEB, 2018).

2. METHODOLOGY

The research was conducted with a descriptive approach. Since the aim of the research is to present an existing situation, the study is descriptive in nature.

2.1. Participants

The study group consisted of students with SLD (n = 994) and students without disabilities (n = 1000) who each sat the 2017-2018 *Central Exam for Secondary Education.* The students without disabilities (n = 1000) were randomly selected from a population of 4986 students. All of the participant students were in their eighth grade, which is the final year of middle school education in Turkey. Students who also had other disabilities (e.g., visual impairment, intellectual disability, etc.) in addition to SLD were excluded from the study (n = 37). In addition, those students who had a course exemption, who had not taken the exams of certain courses, or who had taken a foreign language exam other than English were also excluded from the study. The selected students with SLD (n = 994) had each taken the same central exam with extended time accommodation. The characteristics of the study group are presented in Table 1.

Table 1. Descriptives of Study Group

			Students without learning difficulties	Students with learning difficulties	Total
School type	Religious School	*f*	111	91	202
		%	5.6	4.6	10.1
	State Middle School	*f*	790	858	1,648
		%	39.6	43.0	82.6
	Private Middle School	*f*	83	43	126
		%	4.2	2.2	6.3
	State Boarding Middle School	*f*	16	2	18
		%	0.8	0.1	0.9
Gender	Male	*f*	505	640	1,145
		%	25.3	32.1	57.4
	Female	*f*	495	354	849
		%	24.8	17.8	42.6
Total		*f*	**1,000**	**994**	**1,994**
		%	**50.2**	**49.8**	**100.0**

According to Table 1, students who sat the exam were mainly educated in state middle schools (82.6%). Although the male and female ratios were close to each other, the male ratio was slightly higher at 57.4%. In line with the research design, the rate of students with SLD and those without disabilities were very close to equal.

2.2. Data Collection

Data were obtained from the Turkish Ministry of National Education's General Directorate of Assessment and Examination Services, following official processes for obtaining such data. The *Central Exam for Secondary Education* was implemented by the Turkish Ministry of National Education for the first time in 2018, based on the eighth-grade curriculum. Consisting of 90 multiple-choice items, the exam is presented in two parts, verbal and quantitative, which are administered in two separate sessions. The verbal domain consists of a total of 50 items in four subtests; Turkish Language (n = 20), Religious Culture and Moral Knowledge (n = 10), Revolutionary History of the Republic of Turkey and Kemalism

(n = 10), and Foreign Language (English) (n = 10). The quantitative domain consists of a total of 40 items in two subtests; Mathematics (n = 20), and Science (n = 20). The duration of the exam's verbal domain is 75 minutes, whilst for the quantitative domain is 60 minutes (MEB,2018). Students with SLD sit the same exam with an additional 20 minutes allowance. In addition, students with disabilities can also request readers and coders (MEB, 2018).

2.3. Data Analyses

Item-based and subtest-based analyses were carried out in order to examine the measurement invariance of the exam, Students with SLD formed the focus group of the study. For an item-based review, the DIF was examined for each subtest. DIF can be observed as consistently favoring one group across the entire ability distribution (uniform), or up to a certain skill level, with one group favoring the other group after a certain skill level (non-uniform) (Swaminathan & Rogers, 1990).

Various techniques exist for DIF analysis (French & Miller, 1996; Svetina et al., 2017; Zumbo, 1999). The current study employed the Mantel Haenszel method based on Classical Test Theory and Lord's chi-square method based on Item Response Theory (IRT). The Mantel-Haenszel is based on the χ^2 statistic which is used to determine uniform DIF. The ΔMH (D) statistic is used to evaluate the DIF level. The DIF size is interpreted according to the absolute value of this statistic. According to the classification of Educational Test Service (Zieky, 2003), Category A is where DIF is absent or negligible ($|D| < 1$), Category B is where there is moderate DIF ($1 < |D| < 1.5$), and Category C is where there is a high level of DIF ($|D| \geq 1.5$).

The literature emphasizes that IRT is offered for DIF due to the invariance of parameters (Kamata & Vaughn, 2004). Lord's chi-square method, which is used in DIF analysis, is also a technique based on IRT that is preferred for determining uniform and non-uniform DIF (Lord, 1980). Since Lord's chi-square is based on IRT, the IRT assumptions were tested. Yen (1993)'s Q3 index was used for the analysis of local independence. The Q3 values calculated for the items were found to be less than .20, which indicates that local independence was achieved (Demars, 2010). Unidimensionality of each subtest was then examined using Modified Parallel Analysis, and scree plots were examined as another evidence for one-dimensionality. As an example, the scree plot of the Turkish Language and Mathematics subtests are presented in Figure 1.

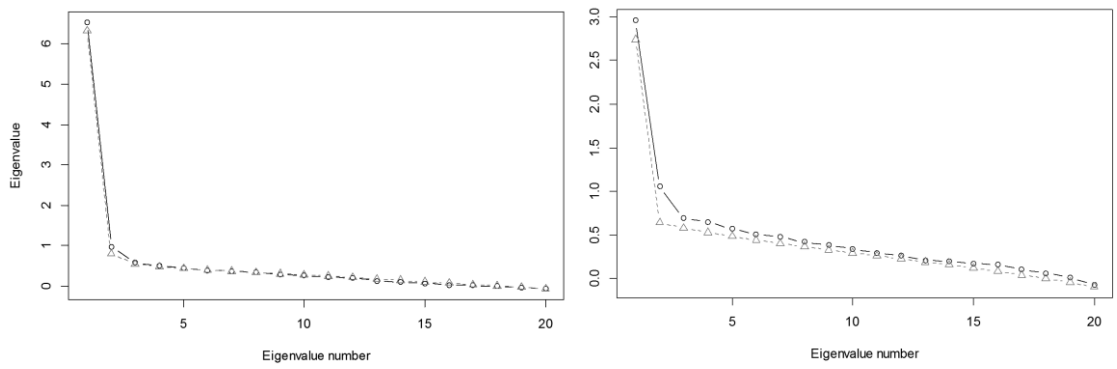

Figure 1. Scree-plots of Turkish and Mathematics Subtests

Figure 1 confirms the presence of a dominant factor in the subtests, and is evidence for the ability to work with one-dimensional models. Model data fit was evaluated using log-likelihood, Akaike Information Criterion and also Bayesian Information Criterion, and it was observed that model fit was higher in the 3PL model. For this reason, analyses related to Lord's chi-square were applied according to the 3PL model.

Multigroup Confirmatory Factor Analysis (MCFA), one of the techniques based on Structural Equation Modeling (SEM), was applied to examine measurement invariance on the basis of each subtest. This technique is one of the leading methods used to measure invariance between groups (Alatlı & Bökeoğlu, 2018; Fischer & Karl, 2019; Van De Schoot et al., 2015). Measurement invariance was investigated in stages with structural, weak, strong, and rigid invariance steps, and the analyses were performed in R with *ltm*(Rizopoulos, 2006), *difR*(Magis et al., 2010), and *lavaan* (Rosseel, 2012) packages.

2.4. Assumption

The number of students with SLD who sat the central exam was within the limit considered acceptable for analysis. In order not to reduce the data volume, item data obtained from different exam test booklets were evaluated together. The item order effect was ignored as the items were aligned according to the different ordering of the different test booklets.

3. FINDINGS AND INTERPRETATION

A summary of the items showing DIF in the *Central Exam for Secondary Education* according to SLD is presented in Table 2.

Table 2. Item Rates Showing DIF in Subtests

	MH			Lord's chi square	Showing DIF in both methods	Showing DIF in at least one method	Showing DIF in favor of reference group in at least one method
	A	**B**	**C**				
Mathematics	4	4	3	19	10	20	8
Turkish Language	5			7	4	8	7
Science	9	4		18	12	19	10
Foreign Language (English)	3	2	1	8	4	10	5
Revolution History of Turkish Republic & Kemalism	3		1	4	3	5	3
Religious Culture & Moral Knowledge	1	1		4	1	5	3
Total	25	11	5	60	34	67	36
%	**27.78**	**12.22**	**5.56**	**66.67**	**37.78**	**74.44**	**40.00**

As summarized in Table 2, 12 items in the Science subtest, 10 items in the Mathematics subtest, four items in the Turkish Language subtest, four items in the Foreign Language (English) subtest, three items in the Revolutionary History of the Turkish Republic & Kemalism subtest, and one item in the Religious Culture & Moral Knowledge subtest showed DIF according to both of the methods applied. Accordingly, a total of 34 items of the 90-item test showed DIF according to both methods. The overall DIF rate of 37.78% is a remarkable finding for a central exam. Among the subtests, those containing the most DIF items were Science and Mathematics, respectively. According to the Mantel Haenszel method, the number of items showing "moderate DIF" or "high DIF" was found to be 16. Considering the number of items in the subtests, it can be stated that DIF rates are very high.

The number of items showing DIF in at least one method was 67 (74.44%), which is a sign of validity regarding measurement invariance across the whole test. At the same time, the number of items showing DIF according to at least one method was 36 in favor of the reference group, as in the group of

1,000 students without disabilities. In other words, 40% of the items showed DIF against individuals with SLD.

Measurement invariance was conducted gradually in the examination of configural invariance, metric invariance, scalar invariance, and strict invariance (Meredith, 1993). Following one after another, each step was verified before the next review was conducted. Configural invariance is considered as the basic test-based measurement of invariance, and shows whether or not tests have the same factor structure across the groups. In another words, it establishes whether or not the items in the test measure the same structure among all groups (Vandenberg & Lance, 1998). According to the results of Multigroup Confirmatory Factor Analysis, configural invariance was provided in all subtests. Metric invariance, which is the second-stage examination of measurement invariance, was not found to have been provided for all the exam's subtests. In Table 3, the results of the Multigroup Confirmatory Factor Analysis applied for configural invariance and metric invariance are summarized. Table 3 presents the X^2 difference, which allows for the examination of differences between fit indices and configural and metric invariance models.

Table 3. Analyze Results of Configural Invariance and Metric Invariance

	Invariance type	**CFI**	**TLI**	**SRMR**	**RMSEA**	**GFI**	**X^2 difference**
Mathematics	Configural	.876	.861	.03	.022	.981	
	Metric	.841	.832	.04	.025	.978	65.461*
Turkish Language	Configural	.949	.943	.03	.025	.988	
	Metric	.897	.890	.05	.034	.98	234.07*
Science	Configural	.919	.910	.03	.027	.982	
	Metric	.903	.897	.04	.029	.979	68.413*
Foreign Language	Configural	.942	.926	.03	.045	.988	
(English)	Metric	.924	.913	.04	.049	.984	54.374*
Revolutionary History	Configural	.961	.950	.03	.035	.995	
of the Turkish Republic	Metric	.923	.912	.05	.047	.992	93.888*
& Kemalism							
Religious Culture &	Configural	.944	.928	.03	.055	.991	
Moral Knowledge	Metric	.912	.899	.06	.065	.986	133.47*

***p < .01**

As can be seen from Table 3, evidence for configural invariance was obtained, which is basic level of measurement invariance. This indicates that the items of the measurement tool represent the

same psychological structure for both individuals with and without SLD. The difference (X^2 difference) between the chi-square values of the structural model and metric invariance model are shown in the rightmost column of Table 3, according to the statistical significance of $p < .01$. This means that metric invariance was not satisfied in any of the subtests. In other words, factor loadings varied between groups in all subtests of the exam. Factor loads of the subtests obtained from the students with SLD and students without disabilities were found to differ. However, since metric invariance was not provided, neither scale invariance or strict invariance was provided.

These results can be interpreted as individuals with and without SLD not responding to the test items in the same way, and thus any comparison of the test scores obtained from different groups cannot be considered meaningful (Steenkamp & Baumgartner, 1998). The lack of measurement invariance is not due to the mean of the measured latent constructs of the group differences in item responses. Since strict invariance was not provided, it may be concluded that the error variances of the responses to the test items are not equal/invariant between the comparison groups.

4. DISCUSSION AND CONCLUSION

According to the findings of this study, measurement invariance was not provided, as either item-based or subtest-based, for the 2017-2018 Turkish Central Exam for Secondary Education. The study's results proved that there was at least one item from each subtest, and a total of 16 items, that showed moderate DIF or high DIF. The subtests containing the most DIF items were identified as Science and Mathematics. Five items were found as showing a high level of DIF, with three from the Mathematics subtest. The number of items showing DIF in the current study were shown to be slightly higher than those previously reported in the literature. In Kamata and Vaughn's (2004) 40-item math test, it was observed that three items showed DIF against individuals with SLD. Similarly, Anjorin (2009) identified items showing DIF worked in favor of individuals without SLD in a high-risk math test.

Displaying DIF for multiple items may also indicate that the test measured different constructs in the group that received the test with accommodations (with SLD) compared to the reference group (without any disability) (Bolt & Ysseldyke, 2008; Kauffman & Hallahan, 2011). Considering this, measurement invariance analyses of the Central Exam for Secondary Education were also conducted on the basis of subject-level subtests. According to the study's findings, configural invariance was provided for all of the subtests, but metric invariance failed. Accordingly, this presents evidence for the subtests measured the same construct for students with and without SLD. However, since metric invariance was not provided, students with and without SLD did not respond to the items in the same way, and thus the comparison of scores obtained from different groups cannot be said to be meaningful (Steenkamp & Baumgartner, 1998). Although the test-takers were at the same ability level, the answers varied because they were in different groups.

In the literature, similar measurement invariance studies have been applied in different large-scale tests (Cook et al., 2010; Kim et al., 2009; Randall & Engelhard, 2010; Steinberg et al., 2011). However, much of the literature conducted in fields such as science and foreign language (English) and at different school levels (e.g., primary and secondary schooling) reported findings that proved that measurement invariance was ensured. Findings that indicate measurement invariance have mostly come from research in which the focus was on whether or not test accommodations were used, and where the studies examined measurement invariance differences along with the use of test accommodations (Randall et al., 2011). In this respect, the lack of metric invariance is a warning for the related test, although the current study measured the same structure between groups in its findings.

Failing to provide adequate measurement and including numerous DIF items does not mean that a test is actually biased towards a certain group (Zieky, 2003); therefore, it may be more appropriate to interpret an item with DIF as being a "probably biased item" (Kamata & Vaughn, 2004). Şenel (2021) also conducted an expert-based bias analysis for the DIF items in the same test, depending on whether they were visually impaired or not. While 16 items were found to show DIF, five items were identified as biased in favor of individuals without a disability, according to expert opinion. Similarly, expert opinion can be sought to conduct bias analysis for items showing DIF.

In Turkey, the score of *Central Exam for Secondary Education*, is the only factor in placement in qualified secondary education institutions that accept students by exam (Milli Eğitim Bakanlığı [Turkish Ministry of National Education, 2018). The significant importance of the central exam scores in the decision process to enter qualified high schools emphasizes the necessity of fair measurement. However, the results of the current study have shown that the central exam does not provide measurement invariance for students with and without SLD. This indicates that comparing students with and without SLD on the basis of this central exam score cannot be taken as meaningful (Steenkamp & Baumgartner, 1998). In other words, relying upon the scores from this centralized test for high school placement may in fact produce unfair results. In conclusion, test developers and designers must consider items showing DIF, especially in high-stake tests, and should work to decrease the number of DIF items. For this aim, it is of vital importance that special education experts are invited to take part in test development, where measures should be taken towards immutability and impartiality, and studies conducted accordingly. In addition, making statistical analyses based on measurement invariance considering relevant disability groups is a requirement for validity evidence of test that include disability groups (American Federation of Teachers et al., 1990). In particular, such analyses are of considerable priority for students with SLD, who account for the highest proportion of students with disabilities.

Students who are both gifted and also have SLD are also frequently observed. In other words, some gifted individuals who show significant potential in certain fields may also experience SLD (Brody & Mills, 1997; Silverman, 2009). The duty of educators is to guide all students to reach their potential. It is also important for students with disabilities to be placed in qualified secondary education schools in order to realize their full potential.

The literature offers that if the performance of students with disabilities in tests is not taken into account, and if the validity of the test scores obtained from these children is not questioned, schools will make less effort for development and achievement of these students (Bolt & Thurlow, 2007). Educational measurement and training has a dynamic and mutually influential relationship. In this respect, ensuring the validity evidence of test applied to students with disabilities, and especially students with SLD, affects the entire education system. Responsiveness in high-stake testing and education as a whole should also be evaluated in the context of human rights.

For practitioners, it is recommended to primarily examine the measurement invariance of large-scale tests in terms of special needs groups. It is also important in terms of accountability to report the bias results in the final reports on large-scale tests. Considering the high rate of individuals with SLD among individuals with special needs, these groups should be given priority. Due to the limited number of studies in this direction, especially in Turkey, it is recommended to conduct research on how fair and valid the measurement and results of these special groups are. Research should be conducted in terms of various dimensions such as measurement invariance of various large-scale tests according to SLD, the effectiveness of the test accommodations used, and the opinions of test takers about the fairness of the tests.

REFERENCES

Abedi, J., Leon, S., & Kao, J. C. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities.* Partnership for Accessible Reading Assessment. https://ici.umn.edu/products/395

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing: National council on measurement in education*. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Alatlı, B. K., & Bökeoğlu, Ö. Ç. (2018). Investigation of measurement invariance of literacy tests in the programme for international student assessment (PISA-2012). *Elementary Education Online*, *17*(2), 1096–1115. https://doi.org/10.17051/ilkonline.2018.419357

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students*.

Anjorin, I. (2009). *High-stakes tests for students with specific learning disabilities: disability-based differential item functioning* [Doctoral dissertation, Southern Illinois University]. https://www.proquest.com/openview/2b3d3f7dd8718df22abe293373d97c35/1?pq-origsite=gscholar&cbl=18750

Bolt, S. E. (2004, April 13). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities* [Conference presentation]. Annual conference of the National Council on Measurement in Education, San Diego, CA.

Bolt, S. E., & Thurlow, M. L. (2007). Item-level effects of the read-aloud accommodation for students with reading disabilities. *Assessment for Effective Intervention, 33*(1), 15–28. https://doi.org/10.1177/15345084070330010301

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment, 26*(2), 121–138. https://doi.org/10.1177/0734282907307703

Borsboom, D. (2006). When does measurement invariance matter?. *Medical Care, 44(*11), S176-S181. doi:10.1097/01.mlr.0000245143.08679.cc

Brody, L. E., & Mills, C. J. (1997). Gifted children with learning disabilities: A review of the issues. *Journal of Learning Disabilities, 30*(3), 282–296. https://doi.org/10.1177/002221949703000304

Brumfield, G. A. (2014). *The effectiveness of reading accommodations for high school students with reading disabilities* [Doctoral dissertation, Walden University]. https://www.proquest.com/openview/8aee69058d23d0cbd915233b60a3a16c/1?pq-origsite=gscholar&cbl=18750

Buzick, H., & Stone, E. (2011). Recommendations for conducting differential item functioning (DIF) analyses for students with disabilities based on previous DIF studies. *ETS Research Report Series, 2011*(2), Article i-26. https://doi.org/10.1002/j.2333-8504.2011.tb02270.x

Büttner, G., & Hasselhorn, M. (2011). Learning disabilities: Debates on definitions, causes, subtypes, and responses. *International Journal of Disability, Development and Education, 58*(1), 75-87.

Camara, W. J., Copeland, T., & Rothschild, B. (2005). Effects of extended time on the SAT ® I: reasoning test score growth for students with learning disabilities. The College Board.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-language arts assessment. *Applied Measurement in Education ISSN*, *23*(2), 187–208. https://doi.org/10.1080/08957341003673831

Cortiella, C. (2005). No Child Left Behind: Determining appropriate assessment accommodations for students with disabilities. National Center for Learning Disabilities.

Demars, C. (2010). Item Response Theory, understanding statistics. Oxford University Press.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*(4), 218–229. https://doi.org/10.1177/00224669070400040301

Elbaum, B., Arguelles, M. E., Campbell, Y., & Saleh, M. B. (2004). Effects of a student-reads-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality*, *12*(2), 71–87. https://doi.org/10.1207/s15327035ex1202_2

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (Eds.). (2018). *Handbook of accessible instruction and testing practices: Issues, Innovations, and Applications* (2nd ed.). Springer.

https://doi.org/10.1007/978-3-319-71126-3

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565–582. https://doi.org/10.1177/0013164406296975

First, M. B. (2013). *DSM-5 handbook of differential diagnosis*. American Psychiatric Publishing.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10,* Article 1507. https://doi.org/10.3389/FPSYG.2019.01507

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities: *Exceptional Children, 72*(2), 136–150. https://doi.org/10.1177/001440290607200201

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities, 45*(2), 128–138. https://doi.org/10.1177/0022219409355484

Grigorenko, E. L., Compton, D. L., Fuchs, L. S., Wagner, R. K., Willcutt, E. G., & Fletcher, J. M. (2019). Understanding, educating, and supporting children with specific learning disabilities: 50 years of science and practice. *American Psychologist, 75*(1), 37-51. https://doi.org/10.1037/AMP0000452

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(2), 49–69.

Kauffman, J. M., & Hallahan, D. P. (Eds.). (2011). *Handbook of special education* (1st ed.). Routledge. https://doi.org/10.4324/9780203837306.ch32

Kavale, K. A., & Forness, S. R. (2000). What definitions of learning disability say and don't say: A critical analysis. *Journal of Learning Disabilities, 33*(3), 239-256.

Kim, D.-H., Schneider, C., & Siskind, T. (2009). Examining the underlying factor structure of a statewide science test under oral and standard administrations: *Journal of Psychoeducational Assessment, 27*(4), 323–333. https://doi.org/10.1177/0734282908328632

Kingsbury, G. G., & Houser, R. L. (1988, April 9). A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing Portland (OR) Public Schools [Conference presentation]. Annual Meeting of the American Educational Research Association, New Orleans, LA. http://iacat.org/sites/default/files/biblio/ki88-01.pdf

Kishore, M. T., Maru, R., Seshadri, S. P., Kumar, D., Sagar, J. K. V., Jacob, P., & Murugappan, N. P. (2021). Specific learning disability in the context of current diagnostic systems and policies in India: Implications for assessment and certification. *Asian Journal of Psychiatry, 55,* 102506.

Knickenberg, M., Zurbriggen, C., Venetz, M., Schwab, S., & Gebhardt, M. (2020). Assessing dimensions of inclusion from students' perspective–measurement invariance across students with learning disabilities in different educational settings. *European Journal of Special Needs Education*, *35*(3), 287–302. https://doi.org/10.1080/08856257.2019.1646958

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California. https://cresst.org/wp-content/uploads/TECH431.pdf

Lai, S. A., & Berkeley, S. (2012). High-stakes test accommodations: research and practice. *Learning Disability Quarterly*, *35*(3), 158–169. https://doi.org/10.1177/0731948711433874

Lindstrom, J. H., & Gregg, N. (2007). The role of extended time on the SAT for students with learning disabilities and/or attention-deficit/hyperactivity disorder. *Learning Disabilities Research & Practice*, *22*(2), 85–95. https://doi.org/10.1111/j.1540-5826.2007.00233.x

Lord, F. M. (1980). Applications of Item Response Theory to practical testing problems. Routledge.

Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Meloy, L. L., Deville, C., & Frisbie, D. (2000, April 26). *The effect of a reading accommodation on standardized test scores of learning disabled and non-learning disabled students* [Conference presentation]. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825.

Middleton, K., & Laitusis, C. C. (2007). Examining test items for differential distractor functioning among students with learning disabilities. *ETS Research Report Series*, *2007*(2), Article i-34. https://doi.org/10.1002/j.2333-8504.2007.tb02085.x

Milli Eğitim Bakanlığı. (2018). Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama klavuzu [Application and implementation guide of central exam for secondary education institutions] Ankara, Turkey. http://www.meb.gov.tr/sinavlar/dokumanlar/2018/MERKEZI_SINAV_BASVURU_VE_UYGULAMA_KILAVUZU.pdf

Mori, K., Tominaga, M., Watanabe, Y., & Matsui, M. (1974). A simple synthesis of methyl 10,11-oxido-3,7,11 -trimethy ldodeca-2,4,6- trienoate, an analog of the Cecropia juvenile hormone. *Agricultural and Biological Chemistry*, *38*(8), 1541–1542. https://doi.org/10.1080/00021369.1974.10861371

National Center for Statistics Education. (2021). Students with disabilities. *The Condition of Education*.

https://www2.ed.gov/programs/osepidea/618-data/state-level-data-files/index.html#bcc;

Ozarkan, H. B., Kucam, E., & Demir, E. (2017). Merkezi ortak sınav matematik alt testinde değişen madde fonksiyonunun görme engeli durumuna göre incelenmesi [An investigation of differential item functioning according to the visually handicapped situation for the Central Joint Exam math subtest]. *Current Research in Education*, *3*(1), 24–34.

Randall, J., Cheong, Y. F., & Engelhard, G. (2011). Using explanatory Item Response Theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129–147. https://doi.org/10.1177/0013164410391577

Randall, J., & Engelhard, G. (2010). Using confirmatory factor analysis and the Rasch Model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education*, *23*(3), 286–306. https://doi.org/10.1080/08957347.2010.486289

Rizopoulos, D. (2006). Itm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2014). *A summary of the research on the effects of test accommodations, 2011-2012 (Synthesis Report 94).* University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/onlinepubs/Synthesis94/Synthesis94.pdf

Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2016). *A summary of the research on the effects of test accommodations: 2013-2014 (NCEO Report 402).* University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Report402/default.htm

Rogers, C. M., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2019). *A summary of the research on effects of test accommodations: 2015-2016 (NCEO Report 412)*. University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport412.pdf

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/JSS.V048.I02

Şenel, S. (2021). Assessing measurement invariance of Turkish "Central Examination for Secondary Education Institutions" for visually impaired students. *Educational Assessment, Evaluation and Accountability, 33,* 621-648. https://doi.org/10.1007/s11092-020-09345-5

Silverman, L. K. (2009). The Two-Edged Sword of Compensation: How the Gifted Cope with Learning Disabilities. *Gifted Education International*, *25*(2), 115–130. https://doi.org/10.1177/026142940902500203

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528/0

Steinberg, J., Cline, F., & Sawaki, Y. (2011). Examining the factor structure of a state standards-based science assessment for students with learning disabilities. *ETS Research Report Series, 2011*(2),

Article i–49. https://doi.org/10.1002/J.2333-8504.2011.TB02274.X

Stone, E., Cook, L., Cahalan Laitusis, C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-Language Arts Assessment for students who are blind or visually impaired. *Applied Measurement in Education*, *23*(2), 132–152. https://doi.org/10.1080/08957341003673773

Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika*, *44*(2), 313–349. https://doi.org/10.1007/s41237-017-0021-0

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

Vandenberg, R. J. & Lance, C. E. (1998). A summary of the issues underlying measurement equivalence and their implications for interpreting group differences. In: *1998 Research Methods Forum*, 3, 1-10.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, *6*, Article 1064. https://doi.org/10.3389/FPSYG.2015.01064

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yılmaz, G. (2019). Seçme sınavlarının engel durumlarına göre madde yanlılığının incelenmesi [An investigation of item bias for selection exams according to disability situations] [Master's thesis, Hacettepe University, Turkey]. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/bitstream/handle/11655/8917/10277911.pdf?sequence=1&isAllowed=y

Zieky, M. (2003). *A DIF Primer*. Center for Education in Assessment.

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (DIF). National Defense Headquarters.