



| Research Article / Araştırma Makalesi |

An Analysis of Item Bias in the PISA 2018 Reading Understanding and Memorising Strategies Questionnaire

PISA 2018 Okuduğunu Anlama ve Hatırlama Stratejileri Anketindeki Maddelerde Yanlılık İncelemesi¹

Sebahat Gören², Ayfer Sayın³, Selahattin Gelbal⁴

Keywords

1. reading comprehension
2. reading strategies
3. PISA 2018
4. item bias
5. measurement invariance

Anahtar Kelimeler

1. okuduğunu anlama
2. okuma stratejileri
3. PISA 2018
4. madde yanlılığı
5. ölçme değişmezliği

Received/Başvuru Tarihi

22.01.2024

Accepted / Kabul Tarihi

25.04.2024

Abstract

Purpose: The study aims to investigate whether there is a bias among countries in the PISA 2018 application regarding the strategies for understanding and memorising in the reading section.

Design/Methodology/Approach: Six countries participated in this descriptive study: The United States, Australia, Korea, Japan, Turkey and New Zealand. The countries were categorized into four groups based on cultural and main language variables, and measurement invariance was initially tested among these groups. Measurement invariance was tested hierarchically between the stages with Multi-Group CFA (MG-CFA) analysis. Subsequently, using the poly-SIBTEST method, it was determined whether the questionnaire items exhibited Differential Item Functioning (DIF) across the groups.

Findings: For the Turkey-USA comparison, which involves different cultures and main languages, expert opinions were obtained for all items including DIF. After the 20 experts, it was concluded that the DIF in three items was due to the item effect. According to expert opinions, the expectation that students in the Turkish sample find the "concentrate on the parts of the text that are easy to understand" strategy less useful. Similarly, there were no identified problems related to translation, culture, curriculum, or other factors in the "summarizing" and "reading aloud to another person" strategies. Expert opinions indicated that three items were biased. The "quickly reading through the text twice" strategy involved skimming the text before reading and reading it twice during the reading process. In the "discussing the content with another person" strategy, it was emphasized that students could misunderstand the discussion. The results showed that the "underlining important parts" strategy could be biased for students in the Turkish sample due to educational programs.

Highlights: Based on the results of this research, it is recommended to review the standards in the translation processes of not only achievement tests but also questionnaire items in international applications such as PISA, whose results are considered important.

Öz

Çalışmanın amacı: Bu araştırmanın amacı, PISA 2018 uygulamasında yer alan okuduğunu anlama ve hatırlama stratejileri anketindeki maddelerin kültür ve ana dil değişkenleri bağlamında ülkeler arasında yanlılık gösterip göstermediğinin incelenmesidir.

Materyal ve Yöntem: Betimsel araştırma türünde yürütülen çalışmaya ABD, Avustralya, Kore, Japonya, Türkiye ve Yeni Zelanda olmak üzere altı ülke dâhil edilmiştir. Ülkeler kültür ve ana dil değişkenlerine göre dört gruba ayrılarak öncelikle gruplar arasında ölçme değişmezliği test edilmiştir. Çoklu grup DFA yöntemi ile ölçme değişmezliği okuduğunu anlama ve hatırlama stratejileri anketi için sırayla şekilsel, metrik, ölçek ve katı değişmezlik olmak üzere aşamalı olarak gerçekleştirilmiştir. Ardından gruplara göre anket maddelerinin değişen madde fonksiyonu (DMF) gösterip göstermediği poly-SIBTEST yöntemi ile belirlenmiştir. Farklı kültür ve ana dile sahip Türkiye-ABD karşılaştırmasında DMF gösteren tüm anket maddelerine göre uzman görüşü alınmıştır.

Bulgular: 20 uzmanın katıldığı çalışmasında üç maddedeki DMF'nin madde etkisinden kaynaklı olduğu ifade edilmiştir. Uzman görüşlerine göre Türkiye örneklemindeki öğrencilerin "metindeki kolay kısımlara yoğunlaşma" stratejisini daha az kullanışlı bulmasının beklendiği bir durum olduğu belirtilmiştir. Benzer şekilde "özetleme" ve "sesli bir şekilde başkasına okuma" stratejilerinde de çevirden, kültürden, eğitim programından ya da başka unsurlardan kaynaklı bir problem olmadığı ifade edilmiştir. Bununla birlikte uzman görüşüne göre anketteki üç maddenin yanlılık gösterdiği tespit edilmiştir. "Metnin geneli hızlı bir şekilde iki kez okuma" stratejisinin alan yazında okuma öncesinde göz gezdirme stratejisi ile okuma sırasında iki kez okuma stratejilerini aynı anda içerdiğine vurgu yapılmıştır. "Metnin içeriğini başkalarıyla tartışma" stratejisinde öğrencilerin soru-cevap olarak nitelendirdikleri stratejide tartışma eyleminin öğrenciler tarafından yanlış anlaşılabilceği vurgulanmıştır. araştırma sonucunda ayrıca "altını çizerek okuma" stratejisinin de eğitim programlarından kaynaklı olarak Türkiye örneklemindeki öğrencilere yanlılık gösterebileceği belirlenmiştir.

Önemli Vurgular: Bu araştırmanın sonuçlarına göre PISA gibi önemli uluslararası uygulamalarda sadece başarı testleri değil duyuşsal alanla ilgili anket maddelerinin de çeviri süreçlerinde standartlara uygun bir şekilde gözden geçirilmesi önem arz etmektedir.

¹ The part of this study was presented as an oral presentation at 8 th CMEEP held on 21-23 September, 2022 in Ege University-İzmir.

² **Corresponded Author**, Hacettepe University, Education Faculty, Educational Sciences, Ankara, TURKEY; <https://orcid.org/0000-0002-6453-3258>

³ Gazi University, Education Faculty, Educational Sciences, Ankara, TURKEY; <https://orcid.org/0000-0003-1357-5674>

⁴ Hacettepe University, Education Faculty, Educational Sciences, Ankara, TURKEY; <https://orcid.org/0000-0001-5181-7262>

INTRODUCTION

Contemporary education aims to raise individuals who can transform knowledge into daily life skills rather than retaining them. In this understanding, which gains momentum, especially with 21st-century skills, the ability to communicate complexly, and share and use knowledge comes to the fore (Joynes et al., 2019). For this, students need to be competent to discover various types of information, recognize and understand how to access them and think about them. The basis of all these processes is reading skills. Reading skill remains the most important skill of the century, even though digital materials are increasing in a rapidly developing world with technology (OECD, 2019a). Similarly, reading skills in the educational environment directly affect students' success not only in language courses but also in other courses. For example, for students to solve a math problem, they need to read and understand the expressions in the problem. A student with well-developed reading skills can reason by making sense of the information about historical events in a course where historical processes are examined. Studies also show that students who actively participate in reading activities have higher learning potential and higher academic achievement compared to their peers (OECD, 2019b).

The reading skills that students acquire and develop in the educational environment also form the basis of their daily life skills. In daily life, the content and price of products can be read while shopping at the market, the contract forms required for any membership can be read and the membership can be decided accordingly, comments can be read to see the opinions of others about a movie watched. In other words, reading comprehension not only supports various disciplines in the field of education but also serves to effectively solve problems in real-life scenarios (Smith et al., 2000). Drawing attention to the importance of the subject, the results of the International Adult Reading Questionnaire (IALS), which was conducted with a large group of participants in the 15-65 age group, reveal that reading skills are effective in all areas of life. The results show that there are significant relationships between individuals' reading skills their lifelong learning efforts and their employment status (Kirsch et al., 2002). All these results show that reading skills are one of the indispensable elements of the educational environment (Kirsch et al., 2002). For this reason, reading strategies (Altinkaya & Sülükçü, 2018; Karatay, 2009; Kuş & Türkyılmaz, 2010) and studies measuring students' reading skills (Çiftçi & Temizyürek, 2008; Gürsoy & Çeliköz, 2022; Kızgın & Baştuğ, 2020; Sayın & Takıl, 2023; Temizkan & Sallabaş, 2015) come to the forefront to improve reading skills because reading comprehension skills form the basis of large-scale tests as well as education.

In our country, large-scale tests such as High School Entrance Exam (LGS), Higher Education Institutions Exam (YKS), Academic Personnel and Graduate Education Entrance Exam (ALES) and Academic Personnel and Graduate Education Entrance Exam (KPSS) include reading comprehension items and literacy skills expected to be used in other subtests such as mathematics. Similarly, international monitoring practices also include items to determine reading skills. For example, the Progress in International Reading Skills Study (PIRLS) measures the reading habits of 4th-grade primary school students and their reading skills in different text types (Mullis et al., 2023). Similarly, the Program for International Student Assessment (PISA) focuses on the extent to which students can use the knowledge and skills they have acquired in daily life, and reading literacy is one of the sub-areas measured (OECD, 2019a). It is seen that the number of countries participating in large-scale applications applied internationally is increasing (Sayın & Takıl, 2017). In addition to ensuring the comparability of countries' education levels, these exams also provide each country with the opportunity to examine its education system from a critical perspective. In this way, countries aim to improve the quality of education by making necessary revisions and improvements in their education policies (Kutlu & Kumandaş, 2015). One of the main reasons for this is that international practices measure not only student achievement or competencies but also the variables that affect them. The PISA application, which is the subject of this study, includes tests to determine students' proficiency in reading, mathematics and science literacy. In addition, data on students' motivation, opinions about themselves, psychological characteristics related to learning processes, school environments and families are collected through questionnaires (MoNE, 2019). Thus, the scores obtained from the tests are scored in the light of the information obtained from the questionnaires. One of the indices used to interpret the reading comprehension scores of 15-year-old students in the PISA application is the UNDREM. To evaluate students' meta-cognitive strategies in PISA, students are given a reading task for each meta-cognitive strategy and they are asked to indicate the reading strategies they use while performing the reading task. In line with the criteria determined based on expert ratings, students' choices are given points and then these scores are standardized and the UNDREM index is produced (OECD, 2019). Within the scope of this study, it was aimed to examine the validity of the answers given to the Reading Understanding and Memorising Strategies Questionnaire Items (ST164), which include reading strategies, in the context of item bias. Considering that validity problems arise especially due to the translation of both test and questionnaire items according to the main language of the countries (Gök et al., 2014; Güner et al., 2014), this study examined the validity evidence for the questionnaire items. The study aimed to determine whether the questionnaire items showed item bias in the Turkish and USA samples with different language and cultural characteristics.

Item bias is when an item in a test gives an advantage or disadvantage to one of the subgroups to which the test is administered, regardless of the trait it is intended to measure (Zumbo, 1999). In the Test Development Standards published by AERA, APA, & NCME in 2014, fairness is defined as the third quality that should be present in tests besides validity and reliability. It means that items should be developed and administered and the results used without giving advantage to any subgroup to ensure equality of opportunity and fairness (Stone-Romero et al., 2010). Otherwise, students will be advantaged or disadvantaged just because they belong to a certain subgroup - race, ethnicity, gender, age, socioeconomic status, cultural structure. Since this

situation affects the scores obtained from the tests, it also negatively affects the decisions made based on the results and therefore needs to be examined (Zumbo, 1999). Item bias is determined in two stages: In the first stage, the differential item function (DIF) is examined. DIF can be defined as the difference in the probability of answering an item correctly among individuals who are similar in terms of the trait measured by the test but who are in different subgroups in terms of variables such as gender, socioeconomic level, and regional differences (Hambleton et al., 1991). After the DIF determination study, content analysis or expert opinion is used to determine item bias. Experts give an opinion on whether the difference in items with DIF is due to item effect or bias (Ackerman, 1992; Osterlind, 1983; Zumbo & Gelin, 2005). In this study, it was examined whether the Reading Understanding and Memorising Strategies Questionnaire Items used in the interpretation of students' reading comprehension scores showed bias according to countries. In the literature, there are studies investigating whether the test items in PISA show DIF (Çelik & Özkan, 2020; Gök et al., 2014; Zeybekoglu et al., 2023) and whether they show item bias (Gökentürk, 2021; Uzun & Gelbal, 2017). However, there are a limited number of studies on the bias of questionnaire items used in the interpretation of scores (Gündeğer, 2023; Schulz, 2008; Köse, 2015; Ünsal Özberk & Koç, 2017). This study will both determine the use of reading strategies in countries and reveal whether there is item bias towards these strategies. The problems of the research are as follows:

- Do the items in the PISA 2018 application's questionnaire on reading understanding and memorising strategies exhibit measurement invariance across cultures and main languages?
- Do the items in the PISA 2018 application's questionnaire on reading understanding and memorising strategies exhibit Differential Item Functioning (DIF) across cultures and main languages?
- In the PISA 2018 application's questionnaire on reading understanding and memorising strategies, do the items exhibiting DIF between Turkey and the United States (different main languages and cultures) also show item bias?

METHOD

Research model

The purpose of this study is to determine whether the items in the PISA 2018 reading understanding and memorising strategies questionnaire show item bias according to main language and culture. In this respect, the study is an example of a descriptive research type. Descriptive research is a type of research that aims to define the current situation and to determine the characteristics, distribution, relationships and change of the research subject (Karasar, 2008).

Participants

Countries

This research has two stages. The first stage consists of students from four countries participating in PISA 2018. PISA 2018 was conducted with the participation of 79 countries, 37 of which are OECD members, and more than half a million students (MoNE, 2019). It aimed to examine the reading strategies subject to the research in terms of main language and culture variables. In this direction, six countries participating in PISA 2018 were identified: the United States of America, Australia, Korea, Japan, Japan, Turkey and New Zealand. Four country sets were formed with these countries according to their cultural and main language characteristics. The country sets are as follows:

- Australia and New Zealand (same main language-similar culture)
- Australia and United States (same main language-different culture)
- Japan and Korea (different main language - similar culture)
- Turkey and United States (different main language-different culture)

The number of students participating in the PISA 2018 application within the scope of the research is shown in Table 1.

Table 1. Countries in the study type

Countries	f	%
Australia	11774	29,6
Japan	5837	14,7
Korea	6260	15,7
New Zealand	4674	11,7
Turkey	6678	16,8
United States of America	4558	11,5
Total	39781	100

Experts

The second participant group of this study consisted of experts. The opinions of 20 experts were consulted to determine the item bias showing DIF in the PISA 2018 reading understanding and memorising strategies questionnaire. The demographic features of the experts participating in the study are shown in Table 2.

Table 2. Demographic features of the experts in the study

Features		f	%
Experts	Turkish education	10	50
	Measurement and evaluation	10	50
Gender	Female	13	65
	Male	7	35
Profession	Master's degree	2	10
	PhD student	5	25
	Prof. Dr.	5	25
	Assoc. Prof. Dr.	6	30
	Asst. Prof. Dr.	2	10
Experience in scale development	Less than 1 year	2	10
	1-5 years	4	20
	5-10 years	7	35
	10-15 years	5	25
	15 years and above	2	10

As seen in Table 2, 50% of the 20 experts participating in the study were experts in the field of measurement and evaluation and 50% were experts in the field of Turkish education. While 65% of the experts were female, 35% were male. 10% of the experts are master's degree graduates and 25% are doctoral students. 25% of the experts are professor doctors, 30% are associate professors and 10% are assistant professors. It was stated that 10% of the experts had less than 1 year of experience in scale development, 20% had 1-5 years of experience, 35% had 5-10 years of experience, 25% had 10-15 years of experience, and 10% had 15 years or more of experience.

Data Collection Tool

In the study, student questionnaire data from the PISA 2018 application conducted by the Organization for Economic Cooperation and Development (OECD) were used. The data were obtained from the official website prepared by the OECD for PISA (<http://www.oecd.org/pisa/>).

The PISA Reading Understanding and Memorising Strategies Questionnaire Items (ST164) are used to identify students' meta-cognition strategies. Meta-cognition is defined as the extent to which students are aware of the strategies they use to answer questions in reading comprehension texts and their ability to use them (OECD, 2019). Studies have shown that students who are aware of their meta-cognition strategies have higher reading comprehension scores (Altunkaya & Sülükçü, 2018; Ohtani & Hisasaka, 2018; Tuna, 2016). In this direction, students are asked to indicate the strategies they use in reading tasks in PISA applications.

All items that make up the Reading Understanding and Memorising Strategies subtest are 6-point Likert type (OECD, 2019). In the Turkish form, the items consist of 6-point response ratings such as "1- Not useful at all 6- Very useful". The items in the Reading Understanding and Memorising Strategies subscale were in English in the Australian, American and New Zealand samples; Japanese in the Japanese sample; Korean in the Korean sample; and Turkish in the Turkish sample. The questionnaire items are presented in Figure 1.

		<i>Not useful at all</i>				<i>Very useful</i>	
		(1)	(2)	(3)	(4)	(5)	(6)
ST164Q01IA	I concentrate on the parts of the text that are easy to understand.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆
ST164Q02IA	I quickly read through the text twice.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆
ST164Q03IA	After reading the text, I discuss its content with other people.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆
ST164Q04IA	I underline important parts of the text.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆
ST164Q05IA	I summarise the text in my own words.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆
ST164Q06IA	I read the text aloud to another person.	<input type="checkbox"/> ₀₁	<input type="checkbox"/> ₀₂	<input type="checkbox"/> ₀₃	<input type="checkbox"/> ₀₄	<input type="checkbox"/> ₀₅	<input type="checkbox"/> ₀₆

Figure 1. PISA reading understanding and memorising strategies questionnaire items (ST164)

Within the scope of the research, an expert opinion form was also used in the final stage of the research, the bias study. The expert opinion form consists of three sections. The first section includes information about the demographic characteristics of the experts. The second section includes explanations. In these explanations, (i) what item bias and its possible causes, (ii) what the PISA reading understanding and memorising strategies questionnaire and its purpose of use, (iii) in which countries the items show DIF and the purpose of the study were mentioned. In the third part of the expert opinion form, a table was prepared for the questionnaire items in which the experts could write their opinions with the answer categories so that they could mark whether there was item bias or not. Within the scope of the research, the opinions of the experts were obtained through one-to-one interviews, and item bias and the characteristics of the questionnaire were explained to each expert. Then, the opinions of the experts were collected face-to-face.

Data Analysis

In the research process, data analysis was carried out in four stages. First, the assumptions required for the analysis were checked and the data were made suitable for analysis. Missing data, outliers, univariate and multivariate normality, multicollinearity and co-variance assumptions were examined for the MG-DFA analysis (Kline, 2012; Tabachnick & Fidel, 2007). After examining the assumptions, missing data were assigned using the Expectation-Maximization algorithm (Kline, 2011), and 973 students with multivariate outliers were removed from the data set. It was determined that the assumption of equi-variance was met and there was no multicollinearity problem in the data set. Although the univariate normality assumption was met in the data set, Mardia's test didn't provide. It was determined that the multivariate normality assumption was not met. The fact that the data are ordered categorically can be shown as evidence that they cannot be normally distributed. Robust maximum likelihood estimation (robust MLE or MLR) takes into account the non-normal distribution in the data (Suh, 2015). Accordingly, MLR estimation was performed with Satorra-Bentler (S-B) correction methods.

In the second stage, measurement invariance analyses were conducted to determine whether the items showed DIF or not. Although many studies do not check the construct validity and intergroup equivalence of validity (Gregorich, 2006), measurement invariance is an assumption that should be checked before making intergroup comparisons (Vandenberg & Lance, 2000). For this reason, within the scope of the study, measurement invariance was tested with the multiple group confirmatory factor analysis (MG-CFA) method depending on the culture and main language subgroups. The factor structure of a test developed in a particular culture and language should be equivalent or similar in another culture where the test is applied (Tabachnick & Fidell, 2014). To test this, the four sub-invariance stages of structural, metric, scalar and strict invariance should be checked (Meredith, 1993). The chi-square difference test is highly affected by the sample size. However, since the difference in CFI are independent of model complexity (Cheung & Rensvold, 2002), nested models can give clearer results. The cut-off score of difference CFI was taken as 0.01. In addition to the fit statistics used in the study were RMSEA, SRMR, TLI, and CFI values which are suggested by Hu and Bentler, 1999. (CFI>0.95, TLI>0.95, SRMR<0.06 and RMSEA<0.08). But TLI value greater than 0.90 is supported as acceptable fit (Hu & Bentler, 1999). The analyses in the measurement invariance stage were conducted with the Mplus8 program for the Windows operating system. In the third stage, whether the items in the student questionnaire contain DIF or not was tested with the poly-SIBTEST method. R Studio program was used in the calculations. In the last stage, interviews were conducted with experts to determine item bias. Content analysis was performed in line with the answers obtained from the experts.

FINDINGS

In this section, for the questionnaire consisting of 6 items in the PISA 2018 Reading Understanding and Memorising Strategies Questionnaire Items, firstly, measurement invariance was examined in line with the Australia-USA, Australia-New Zealand, Japan-Korea, Turkey-USA, samples. Then, it was examined whether the items showed DIF based on these country sets. Finally, it was determined whether the items showing DIF in the USA-Turkey comparison showed bias or not.

Research model understanding and memorising strategies exhibit measurement invariance across cultures and main languages?

In the first sub-problem of the study, the answer to the question " Do the items in the PISA 2018 application's questionnaire on reading understanding and memorising strategies exhibit measurement invariance across cultures and main languages? is sought. Measurement invariance was tested in stages with the MGCFA method. Since each invariance stage is a prerequisite for the next stage, structural, metric, scalar and strict invariance tests were carried out respectively. Before proceeding to the measurement invariance stages, CFA was conducted for the single-factor structure of the questionnaire items for the all data and each country. The results of the analysis for all data are shown in Figure 1. It is recommended to conduct CFA for each group before starting measurement invariance studies (Van de Schoot et al., 2012) and this CFA model to be examined should show a good fit for each group. The model-fit indices calculated after the country-based analysis are also shown in Table 3.

Table 3. Model fit indices of the single factor structure of PISA 2018 reading understanding and memorising questionnaire by countries

Countries	χ^2	sd	RMSEA	SRMR	CFI	TLI
America (USA)	28.69	7	0.026	0.028	0.995	0.990
Australia (AUS)	153.62	7	0.042	0.016	0.989	0.977
Japan (JPN)	201.91	8	0.069	0.026	0.971	0.937
Korea (KOR)	247.97	8	0.078	0.031	0.972	0.939
New Zealand (NZL)	127.63	8	0.061	0.022	0.963	0.921
Turkey (TUR)	180.51	8	0.057	0.022	0.971	0.946

df = degrees of freedom, RMSEA= root mean square error of approximation, CFI= comparative fit index, TLI=Tucker-Lewis Index, SRMR = standardized root mean square residual.

When Table 3 is examined, it is seen that the six items in the PISA 2018 Reading Understanding and Memorising Questionnaire show a single-factor structure in the data of the six countries in the study. In all six countries, CFI values were found between 0.953 and 0.971, TLI values between 0.911 and .946, SRMR values between 0.022 and 0.038, and RMSEA values between .052 and .079. This information means that the model-data fit is good in all samples (CFI>0.95, TLI>0.90, SRMR<0.06 and RMSEA<0.08) (Hu & Bentler, 1999). After this section, the measurement invariance stages of the model in all sample pairs were tested respectively. The results are shown in Table 4.

Table 4. Model-data fit indices for measurement invariance

Countries	χ^2	df	RMSEA	SRMR	CFI	TLI	Δ CFI
AUS-NZL							
Configural invariance	225.66	14	0.043	0.016	0.988	0.974	-
Metric invariance	268.38	19	0.040	0.023	0.986	0.977	0.003
Scale invariance	324.46	24	0.039	0.026	0.983	0.978	0.001
Strict Invariance	359.59	30	0.037	0.035	0.981	0.981	0.002
AUS-USA							
Configural invariance	183.48	14	0.039	0.014	0.991	0.980	-
Metric invariance	198.36	19	0.034	0.015	0.990	0.974	0.001
Scale invariance	259.55	24	0.035	0.017	0.987	0.984	0.003
Strict Invariance	418.96	30	0.040	0.031	0.978	0.978	-0.009
JPN-KOR							
Configural invariance	482.67	14	0.074	0.029	0.972	0.940	-
Metric invariance	609.76	19	0.078	0.038	0.965	0.945	0.007
Scale invariance	1982.22	24	0.116	0.069	0.884	0.855	-0.081
Strict Invariance	-	-	-	-	-	-	-
TUR-USA							
Configural invariance	145.10	14	0.041	0.016	0.987	0.973	-
Metric invariance	185.04	19	0.039	0.023	0.984	0.975	0.003
Scale invariance	563.98	24	0.037	0.040	0.948	0.935	-0.036
Strict Invariance	-	-	-	-	-	-	-

df = degrees of freedom, RMSEA= root mean square error of approximation, CFI= comparative fit index, TLI=Tucker-Lewis Index, SRMR = standardized root mean square residual.

The examination of measurement invariance first started with configural invariance. At this stage, it was examined whether the questionnaire has the same general factor structure across countries (whether it shows a single factor structure). In summary, in this stage, which is also called configural invariance, the comparability of the structure across countries was examined. When the fit indices in Table 4 are examined, the fit statistics at the configural invariance stage are within the accepted range (RMSEA<0.08; SRMR<0.06; CFI>0.95; TLI>0.90). Only χ^2 /df values were not taken into consideration since they may give biased results in large samples (Kline, 2016). From this, it can be concluded that the model provides configural invariance across countries. Then, the second stage of measurement invariance, metric invariance, was tested. In the metric invariance stage, factor and item groups in the model as well as item factor loadings were equalized across countries and invariance was tested. Factor loadings were constrained to be equal across groups, but group-specific models were estimated simultaneously by allowing other model parameters to differ across groups. When the fit statistics in Table 4 are examined, it is seen that the model fits the data at the metric invariance stage (RMSEA<0.08; SRMR<0.06; CFI>0.95; TLI>0.90). To determine whether metric invariance was achieved, the difference in CFI values were examined and it was determined that the values were within the desired range (Δ CFI<0.01). Therefore, metric invariance of the model was ensured and the scale invariance phase was started. In the scale invariance stage, in addition to item factor groups and item factor loadings, item constants were also equalized between groups and invariance was tested. Thus, the comparability of the observed variable and factor loading mean values was examined (Gregorich, 2006). At this stage, when Table 4 is examined, it is seen that the model fits the data at the scale invariance stage in the groups other than the groups speaking the same main language (RMSEA<0.08; SRMR<0.06; CFI>0.95; TLI>0.90).). It was determined that the change values in CFI values were within the acceptable range (Δ CFI <0.01) to decide whether scale invariance was achieved. Therefore, scale invariance was ensured between the model groups and the last stage of measurement invariance, which is strict invariance, was tested. In the strict invariance stage, item factor groups, item factor loadings and item constants as well as item residual variances were equalized and invariance was tested (Wu et al., 2007). Since scale invariance was not supported in the different language-same culture samples (JPN-KOR,) and different language-different culture (TUR-USA) strict invariance was tested in all but these groups. When the fit statistics in Table 4 are examined, it is seen that the model fits the data at the strict invariance stage in AUS-NZL, AUS-USA, samples (RMSEA<0.08; SRMR<0.06; CFI>0.95; TLI>0.90).The difference in CFI values (Δ CFI <0.01), which are examined to determine whether strict invariance is achieved, shows that strict invariance is achieved between the model groups. Therefore, it was determined that the questionnaire items showed a similar structure in countries matched with certain variations in terms of culture and same main language. After this determination, we proceeded to the stage of whether the questionnaire items for reading understanding and memorising strategies show DIF according to the countries.

Do the items in the PISA 2018 application's questionnaire on reading understanding and memorising strategies exhibit Differential Item Functioning (DIF) across cultures and main languages?

It was tested whether the items in the Reading Understanding and Memorising Strategies questionnaire in the PISA 2018 application showed DIF between Australia-New Zealand (same main language-similar culture), USA-Australia (same main language-different culture), Japan-Korea (different language-similar culture) and USA-Turkey (different language-different culture) according to the poly-SIBTEST method. The results are shown in Table 5.

Table 5. DIF analysis results of the reading understanding and memorising strategies questionnaire by cultures

Countries	Items	β	Level	Advantage
AUS-NZL (same main language-similar culture)	ST164Q01IA	0.010	-	-
	ST164Q02IA	0.111*	C	AUS
	ST164Q03IA	-0.008*	-	-
	ST164Q04IA	-0.132*	C	NZL
	ST164Q05IA	-0.012	-	-
	ST164Q06IA	-0.098	C	NZL
AUS-USA (same main language-different culture)	ST164Q01IA	0.078*	B	USA
	ST164Q02IA	-0.109*	C	AUS
	ST164Q03IA	0.108*	C	USA
	ST164Q04IA	0.025	-	-
	ST164Q05IA	0.024	-	-
	ST164Q06IA	-0.084*	B	AUS
JPN-KOR (different language-similar culture)	ST164Q01IA	0.139*	C	JPN
	ST164Q02IA	0.796*	C	JPN
	ST164Q03IA	-0.363*	C	KOR
	ST164Q04IA	-0.206*	C	KOR
	ST164Q05IA	0.106*	C	JPN
	ST164Q06IA	0.051*	A	JPN

TUR-USA (different language-different culture)	ST164Q01IA	0.157*	C	USA
	ST164Q02IA	0.067*	B	USA
	ST164Q03IA	0.444*	C	USA
	ST164Q04IA	-0.241*	C	TUR
	ST164Q05IA	-0.122*	C	TUR
	ST164Q06IA	-0.341*	C	TUR

* $p < .05$

According to the criteria proposed by Roussos & Stout (1996), $|\beta| < 0.059$ is negligible (A), $0.059 \leq |\beta| < 0.088$ is moderate (B) and $|\beta| > 0.088$ is high (C). The positive β value obtained indicates that the item shows DIF in favour of the reference group and the negative β value indicates that the item shows DIF in favour of the focus group (Abbott, 2007). In this direction, when the information in Table 5 is analyzed, it is seen that three items show DIF according to Australia and New Zealand, who speak the same main language and have similar cultures. According to the USA and Australia, who speak the same main language but have different cultures, 3 items show DIF. According to Japan and Korea, which have different languages and similar cultures, all 6 items in the questionnaire show DIF. However, it can be said that the DIF of the sixth item is at a negligible level; in fact, five items show DIF. Similarly, between the USA and Turkey, who speak different native languages and have different cultures, six items show DIF and all DIFs are at high levels.

In the PISA 2018 application's questionnaire on reading understanding and memorising strategies, do the items exhibiting DIF between Turkey and the United States (different main languages and cultures) also show item bias?

In the calculations in the PISA 2018 reading understanding and memorising strategies questionnaire based on Turkey and the USA, the USA was determined as the reference group and Turkey was determined as the focus group. Accordingly, it was determined that items 1, 2 and 3 of the questionnaire showed DIF in favour of the USA, and items 4, 5 and 6 showed DIF in favour of Turkey. In other words, the students in the USA sample found the strategies in the first three items of the questionnaire more useful than the students in Turkey. In the last three items of the questionnaire, students in the Turkey sample found the reading strategies more useful than students in the USA sample. To determine whether this result is an item effect or an item bias, the opinions of 20 experts were taken. The frequency and percentage values of the experts' opinions are shown in Table 6.

Table 6. Expert opinions on the bias of the items in the PISA 2018 reading understanding and memorising strategies questionnaire-frequency and percentage values

Advantageous group	Items	Turkish education experts (f)	Measurement and evaluation experts (f)
USA	ST164Q01IA- I concentrate on the parts of the text that are easy to understand.	0	2
	ST164Q02IA- I quickly read through the text twice.	7	2
	ST164Q03IA- After reading the text, I discuss its content with other people.	9	4
Turkey	ST164Q04IA- I underline important parts of the text.	7	3
	ST164Q05IA- I summarise the text in my own words.	3	3
	ST164Q06IA- I read the text aloud to another person.	4	2

When the information in Table 6 is examined, it is seen that the opinions of Turkish education experts and measurement and evaluation experts differ in their views on item bias. More than 70% of Turkish education experts stated that there was bias in 4 items. On the other hand, measurement and evaluation experts generally stated that there was no bias in the items. The opinions of Turkish education experts on the questionnaire items that they stated to have item bias were analyzed by content analysis. The opinions of the experts are as follows:

For the first item of the questionnaire, "I concentrate on the parts of the text that are easy to understand.", none of the Turkish educators (0%) stated that there was no bias in the item. Twenty percent of the measurement and evaluation experts stated that there might be bias in the item. The expert coded OU3 and ME5 stated that the expression "easy part" may be relative. In line with the opinions of the experts, it is thought that the DIF in the first item of the questionnaire is caused by the item effect.

Regarding the second item of the questionnaire, "I quickly read through the text twice." 70% of the Turkish education experts and 20% of the measurement and evaluation experts stated that there was a bias in the item. The opinions of the experts on the subject are as follows. TE1 "Students generally read the texts once and then answer the questions due to the habit they have acquired in the process. For this reason, students in the Turkish sample may have found this strategy impractical, thinking that it is not correct." TE3 "Although the students found the part of reading the text quickly useful, they may have given a low score to this item because of the part of reading the text twice." ME4 "There are two judgments, individuals from different cultures may have focused on this differently." Four other experts expressed similar opinions. In line with the opinions of the experts, it can be said that there is an item bias due to the two actions in the item.

Regarding the third item of the questionnaire, "After reading the text, I discuss its content with other people." 90% of the Turkish education experts and 40% of the measurement and evaluation experts stated that there was a bias in the item. The expert coded TE1 said, "The expression 'discuss' in the item is generally used in a negative sense in Turkish. Therefore, students may have given low scores to this item." Similarly, TE5 said, "I think there is a translation error in the item, the word 'discuss' is wrong." The expert coded OU9 said, "Arguing can mean good or bad in different cultures." All of the experts who stated that there was a bias in the item in question stated that the word "discuss" in this item caused a wrong connotation in the students in the Turkish sample and therefore it was biased. Accordingly, it can be said that there is an item bias in this item due to translation.

Regarding the fourth item of the questionnaire, "I underline important parts of the text.", 70% of the Turkish education experts and 30% of the measurement and evaluation experts stated that there was a bias in the item. Among the experts who expressed their opinions on this subject, an expert coded TE2 said, "In the Turkish education system, children study by underlining from primary school onwards. This strategy is very useful for them." TE4 said, "Our students like to underline words in the text." ME10 said, "Studying by underlining is a strategy that students use a lot." This situation shows that there may be a bias in the item due to the education program.

Regarding the fifth item of the questionnaire, "I summarise the text in my own words.", only 30% of both Turkish language education and Measurement and Evaluation experts stated that there was an item bias in the item. The expert coded TE4, who expressed an opinion on the subject, stated that the expression "using their own words" may confuse students. Similarly, the ME7 stated that the text can be summarized "without their own words". Considering both the DIF in favour of Turkey and the expert opinions on the item, it can be said that there is an item effect in the item.

In the sixth item of the questionnaire, "I read the text aloud to another person.", 40% of the Turkish education experts and 20% of the measurement and evaluation experts stated that there was a bias in the item. The TE3 said, "How will this item be used in solving the reading test, it seems like there is a mistake." TE6 said, "Students may not have understood why the text would be read aloud to someone else if it is not in the classroom environment and the teacher did not ask for it." Although other experts did not express similar opinions on the subject, it was stated that the item should have worked against Turkey. However, this item showed DIF in favour of Turkey. In line with both these opinions and expert opinions, it can be stated that the item did not show bias and that the DIF was due to the item effect.

CONCLUSION, DISCUSSION AND SUGGESTIONS

This study aims to examine whether the items in the reading understanding and memorising strategies questionnaire in the PISA 2018 application show item bias according to culture and main language. In this direction, firstly, the countries including same main language - similar culture; same main language - different culture; different language - similar culture; different language-different culture comparisons were determined. For this purpose, Australia, New Zealand, USA, Japan, Korea and Turkey samples were decided and pairwise comparisons were made. Within the scope of the research, firstly, the factor structure of the questionnaire items was examined and it was determined that the items showed a single-factor structure. The findings revealed that students' Reading Understanding and Memorising Strategies questionnaire items showed a unidimensional structure. It was determined that measurement invariance was achieved in four different variations other than strict invariance. In applications that focus on comparing countries such as PISA, measurement invariance should be examined before comparison (Vandenberg & Lance, 2000). Similarly, there are studies in the literature examining measurement invariance before DIF analyses (Asil & Gelbal, 2012; Karakoç Alatlı & Cokluk Bökeoğlu, 2018; Tiryaki, 2018).

After measurement invariance, it was examined whether there were DIFs in the reading understanding and memorising strategies questionnaire items in PISA 2018. As a result of the research, it was determined that there were fewer DIFs in the questionnaire items in countries with the same main language, and more (all) DIFs in countries with different main languages. In applications such as PISA, there are DIF studies for the answers of students who speak different main languages and belong to different cultures (Gelbal & Uzun, 2017; Grisay et al., 2019; Gür, 2019; Kankaraš & Moors, 2014; Lee, 2009; Yildirim & Berberoglu, 2009). These results indicate that there may be a problem arising from the translation of items into different main languages. In other words, it can be thought that translation problems (also due to cultural differences) continue in multilingual applications such as PISA.

Finally, it was determined whether the questionnaire items showed bias in the Turkey-USA sample according to expert opinions. For this, the items with DIF were first analyzed. Although the perception of finding reading strategies useful after DIF was the same, it was determined that the students in the USA sample found the strategies "I concentrate on the parts of the text that are easy to understand.", "I quickly read through the text twice." and "After reading the text, I discuss its content with other people." are more useful than the students in Turkey (they scored higher on these items). Students in the Turkey sample found the strategy "I underline important parts of the text.", "I read the text aloud to another person." and "I summarise the text in my own words." more useful than students in the USA. Whether this difference (DIF) was due to item bias or item effect was examined according to expert opinion (Zumbo, 1999). For this purpose, the opinions of 10 Turkish education experts and 10 measurement and evaluation experts were taken. As a result of the research, it was determined that the opinions of Turkish education experts and measurement and evaluation experts did not overlap. The majority of measurement and evaluation experts stated that there was no item bias in the questionnaire items. This result may be although measurement and evaluation experts know the technical

concept of item bias better, they do not have sufficient knowledge in areas such as curriculum implementation, cultural differences, and reading strategy use. Since the expert opinions were taken face-to-face in the study and explanations were made about the process, it can be thought that Turkish educators gave more reliable answers in terms of subject matter expertise. Sayin and Sata (2022) also found that Turkish education experts were more reliable than measurement and evaluation experts in scoring the reading test developed by the students. These results point to the importance of expert selection and the need for a combination of subject matter and technical knowledge.

According to the expert opinions, strategies with DIF and item effects were analyzed. It was determined that the students in the Turkey sample found the strategy of concentrating on the easy parts of the text less useful than the students in the USA sample. However, summarizing the text in their own words and reading the text aloud to another person were found to be more useful by the students in the Turkey sample. In the national literature, reading comprehension strategies are classified as strategies used before, during and after reading. Strategies such as goal setting and skimming are used before reading; strategies such as fluent reading, checking the meaning, underlining, and finding the meaning of unfamiliar words are used during reading; and strategies such as summarizing, analyzing, and evaluating are used after reading (Aktas & Bayram, 2018; Demirel & Epeçan, 2012; Temizkan, 2007). The first item of the PISA Reading Understanding and Memorising Questionnaire includes the expression "concentrating on easy-to-understand parts of the text". To improve Turkish vocabulary, teachers and students focus on the strategy of "finding the meaning of unknown words" in reading strategies (Baydik, 2011; Sevim & Söylemez, 2018). Therefore, it can be said that this strategy shows a DIF against Turkey, in other words, students find this strategy less useful than the USA sample. In contrast to this situation, some studies support that students use summarizing and sharing strategies intensively in the classroom environment. For example, Aktas and Bayram (2018) stated in their study that summarizing is the strategy that classroom teachers use the most after reading. Therefore, these strategies may have been high among the students in the Turkish sample due to the habits acquired in the process.

The majority of Turkish education specialists stated that there was bias in three items. The second item of the questionnaire, "I quickly read through the text twice." was the only item that showed DIF in all comparisons. Experts also stated that the fact that the item contains two statements, it is not clear whether it focuses on fast reading or reading twice, and that the item shows bias for this reason. In the statement "After reading the text, I discuss its content with other people.", it was stated by the experts that there might be a translation bias because the use of the word "discuss" in Turkish carries negative connotations. This situation indicates that there are some problems arising from cultural and translation characteristics. Asil and Gelbal (2012) also found that the questionnaire items showed DIF in different native-speaker countries. Similarly, Van de Vijver and Tanzer (2004) stated that the possible causes of DIF could be inadequate translations, ambiguity in the meaning of the original item, familiarity with the item content in certain cultures, or culture-specific implications related to the way the item is written. In this study, skim reading is a strategy used before reading, while second reading is a strategy used during reading (Aktas & Bayram, 2018; Yilmaz, 2008). Similarly, the word "discuss" may refer to the question-answer strategy, but it is possible to talk about a linguistic difference here.

Based on the results of this research, it is recommended to review the standards in the translation processes of not only achievement tests but also questionnaire items in international applications such as PISA, whose results are considered important. Considering the effect of reading strategies on achievement, it is recommended to compare the reading strategies in PISA questionnaires with the reading strategies applied in the educational environment. Instead of poly-SIBTEST, which was used in the process of determining DIF in this study, other methods or different subgroups can be determined and the research can be repeated. In the process of determining item bias, people who are experts in different fields can be brought together and their opinions can be obtained through applications such as the Delphi technique, panel interviews, etc.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, author-ship, and/or publication of this article.

Statements of publication ethics

We hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers' contribution rate

The study was conducted and reported with equal collaboration of the researchers.

Ethics Committee Approval Information

This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

REFERENCES

- Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, 24(1), 7-36. <https://doi.org/10.1177/0265532207071510>
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67-91. <https://doi.org/10.1111/j.1745-3984.1992.tb00368.x>
- Aktas, E., & Bayram, B. (2018). Türkçe öğretiminde okudugunu anlama stratejilerinin kullanımı üzerine bir inceleme. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 22(3), 1401-1414.
- Altunkaya, H., & Sülükçü, Y. (2018). The Correlation Between Reading Strategy Metacognitive Awareness Levels and Reading Comprehension Skills of 7th Grade Students. *Uluslararası Türkçe Edebiyat Kültür Eğitim (TEKE) Dergisi*, 7(4), 2502-2517.
- Asar, E. (2019). *An investigation of the measurement of the measurement between speaking countries of different languages of the PISA 2015 mathematics literacy test.* (Thesis No. 551108) [Master's thesis, Akdeniz University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Asil, M. & Gelbal, S. (2012). Cross-cultural Equivalence of the PISA Student Questionnaire. *Education and Science*, 37(166), 236-249.
- Batur, Z., Ulutas, M., & Beyret, T. N. (2019). Investigation of 2018 LGS Turkish questions for PISA reading skills objectives. *Milli Eğitim Dergisi*, 48(1), 595-615.
- Baydik, B. (2011). Examining the use of metacognitive reading strategies of students with reading difficulties and their teachers' reading comprehension instruction practices. *Education and Science*, 36(162).
- Çelik, M., & Özkan, Y. Ö. (2020). Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions *Journal of Measurement and Evaluation in Education and Psychology*, 11(3), 283-301. <https://doi.org/10.21031/epod.715020>
- Çiftçi, Ö. & Temizyürek, F. (2008). Measurement understanding of reading skills in 5th classes of primary schools. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 5(9), 109-129.
- Demirel, Ö. ve Epçaçan, C. (2012). Effects of reading comprehension strategies on cognitive and affective learning outcomes. *Kalem Eğitim ve İnsan Bilimleri Dergisi*, 2(1), 71-106.
- Gök, B., Kabasakal, K. A., & Kelecioğlu, H. (2014). Analysis of attitude items in PISA 2009 Student Questionnaire in terms of differential item functioning based on culture. *Journal of Measurement and Evaluation in Education and Psychology*, 5(1), 72-87. <https://doi.org/10.21031/epod.64124>
- Göktentürk, T. (2021). Dif analysis of the pisa reading test according to the ecological model: The example of Turkey (Thesis No. 680538) [Doctoral dissertation, Yildiz Teknik University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44(11), 78-94. <https://doi.org/10.1097/01.mlr.0000245454.12228.8f>
- Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI*, 2.
- Gülleroğlu, H. D. (2017). An investigation of measurement invariance by gender for the Turkish students' affective characteristics who took the PISA 2012 math test. *GÜJGEF*, 37(1), 151-175.
- Gündeger, C. (2023). An Investigation of Item Bias for the Attitude Scale Towards E-Assessment. *Bolu Abant İzzet Baysal University Journal of Faculty of Education*, 23(1), 249-271. <https://dx.doi.org/10.17240/aibuefd.2023-1192209>
- Güner, H., Çelebi, N., KAYA, G. T., & Korumaz, M. (2014). Analysis of international tests (PISA, TIMSS and PIRLS) in the context of neoliberal education policies and equality of opportunity in education. *Journal of History Culture and Art Research*, 3(3), 33-75. <https://doi.org/10.7596/taksad.v3i3.329>
- Gür, E. (2019). *An investigation of the PISA 2015 in terms of differential item functioning based on culture.* (Thesis No. 582227). [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Gürsoy, S. E., & Çeliköz, N. (2022). The effect of primary school 2nd grade students' reading comprehension skills in Turkish classes on the problem-solving skills in math classes. *International Journal of Social and Educational Sciences*, 17, 78-96. <https://doi.org/10.20860/ijoses.1084867>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications, Inc.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Imrol, F. (2017). *Investigation of measurement invariance of motivation and self-belief constructs towards mathematics in PISA 2012 Turkey sample* (Thesis No. 456693) [Master's thesis, Akdeniz University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Joyes, C., Rossignoli, S., & Fenyiwa Amonoo-Kuofi, E. (2019). 21st Century Skills: Evidence of issues in definition, demand and delivery for development contexts (*K4D Helpdesk Report*). Brighton, UK: Institute of Development Studies.
- Kaman, S. (2012). *Effect of using reading fluency strategies on improving reading skills of third grade students at primary schools* (Tez No. 300682). [Master's thesis, Ahi Evran University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Kankaraş, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, 45(3), 381-399.
- Karakoç Alatlı, B. & Çokluk Bökeoğlu, Ö. (2018). Uluslararası Öğrenci Değerlendirme Programı (PISA-2012) okuryazarlık testlerinin ölçme degismezliginin incelenmesi. *İlköğretimOnline*, 17(2), 1096-1115. <https://doi.org/10.17051/ilkonline.2018.419357>
- Karatay, H. (2009). Okuma stratejileri bilisel farkındalık ölçeği. *Bolu Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 9(2), 58-79. <https://doi.org/10.11616/AbantSbe.242>
- Kibrislioglu, N. (2015). *The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey - China (Shangai) - Indonesia* (Tez No. 394817). [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>
- Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). Reading for change: Performance and engagement across countries: Results from PISA 2000. <https://search.informit.org/doi/10.3316/aeipt.123439>

- Kizgin, A., & Bastug, M. (2020). Predicting level of reading motivation and reading comprehension skill on academic success. *Journal of Language Education and Research*, 6(2), 601-612. <https://doi.org/10.31464/jlere.767022>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford.
- Köse, I. A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (q32-q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi. *Kastamonu Eğitim Dergisi*, 23(1), 227-240.
- Kus, Z. & Türkyılmaz, M. (2010). Sosyal bilgiler ve Türkçe öğretmeni adaylarının okuma durumları: ilgi, alışkanlık ve okuma stratejilerini kullanım düzeyleri. *Türk Kütüphaneciliği*, 24(1), 11-32.
- Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy and math anxiety across 41 countries in PISA 2003 participating countries. *Learning and Individual Differences*, 19, 255-365. <https://doi.org/10.1016/j.lindif.2008.10.009>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <https://doi.org/10.1007/BF02294825>
- Millî Eğitim Bakanlığı (2019). *PISA 2018 Turkey Raporu*. MEB Yayınları.
- OECD (2019a). "PISA 2018 Reading Framework", in *PISA 2018 Assessment and analytical framework*, OECD Publishing, Paris, <https://doi.org/10.1787/5c07e4f1-en>.
- OECD (2019b). *The questionnaire of Adult Skills : Reader's Companion, Third Edition*, OECD Skills Studies, OECD Publishing, Paris, <https://doi.org/10.1787/f70238c7-en>
- Sayın, A., & Takil, N. B. (2023). Difficulties Encountered In Preparing Items For Measuring High-Level Thinking Skills And Suggestions For Solutions: Reading Comprehension. *GEFAD*, 43(1), 531-561. <https://doi.org/10.17152/gefad.1195081>
- Schulz, W. (2008). *Questionnaire construct validation in the international civic and citizens hipeducation study*. Paper presented to the 3rd IEA International Research Conference in Taipei, September.
- Sevim, O., & Söylemez, Y. (2018). Okuma eğitimi. *Nobel Yayincılık*.
- Smith, M. et al. (2000), "What Will Be the Demands of Literacy in the Workplace in the Next Millennium?", *Reading Research Quarterly*, Vol. 35(3), pp. 378-383, <http://dx.doi.org/10.1598/RRQ.35.3.3>
- Spiro, R. J., DeSchryver, M., Hagerman, M. S., Morsink, P. M., & Thompson, P. (Eds.). (2015). *Reading at a crossroads?: Disjunctures and continuities in current conceptions and practices*. Routledge.
- Stone-Romero, E. F., & Rosopa, P. J. (2010). Research design options for testing mediation models and their implications for facets of validity. *Journal of Managerial Psychology*, 25(7), 697-71.
- Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling*, 22(4), 568-80. <https://doi.org/10.1080/10705511.2014.937669>
- Tabachnick, B. G. & Fidell, L.S. (2007). *Using multivariate statistics*. Pearson Education.
- Takil, N. B., & Sayın, A. (2017). Opinions of the Turkish teacher candidates for change in the reading skills of the students in the 15-year-old group. *International Journal of Language Academy*, 5(15), 266-284. <https://doi.org/10.18033/ijla.3561>
- Temizkan, M., & Sallabaş, M. E. (2015). Okudugunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, (30), 207-220.
- Tiryaki, F. (2019). *Analysis on differential item functioning and measurement invariance of PISA 2015 student attitude surveys* (Tez No. 558074). [Master's thesis, Ankara University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>.
- Tuna, L. (2016). *The effect of reading strategy use of 7th graders on reading habit acquisition, attitude and comprehension skills* (Thesis No. 430703). [Master's thesis, Dokuz Eylül University]. Council of Higher Education Thesis Center. <https://tez.yok.gov.tr>.
- Ünsal Özberk, E. B., & Koç, N. (2017). Gender and socioeconomic status DIF on the wis-civ Turkish form items: a comparison of dif detection techniques. *Journal of Measurement and Evaluation in Education and Psychology*, 8(1), 112-127. <https://doi.org/10.21031/epod.287577>
- Uzun, N. B., & Gelbal, S. (2017). An investigation of item bias in pisa science test in terms of the language and culture. *Kastamonu Eğitim Dergisi*, 25(6), 2427-2446.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European journal of developmental psychology*, 9(4), 486-492.
- Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, 54(2), 119-135. <https://doi.org/10.1016/j.erap.2003.12.004>
- Vanderberg, R. J. & Lance, C. E., (2000). A Review and synthesis of the measurement invariance literature: suggestions practices, and recommendations for organizational research. *Organizational Research Methods*. 3(4), 4-70. <https://doi.org/10.1177/109442810031002>
- Wu, A. D., Li, Z., & Zumbo, B. D. (2019). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation*, 12(1), 3. <https://doi.org/10.7275/mhqa-cd89>
- Yildirim, H. H. & Berberoglu, G. (2009). Judgemental and statistical analyses of the PISA 2003 mathematics literacy items. *International Journal of Testing*, 9(2), 108-121. <https://doi.org/10.1080/15305050902880736>
- Yılmaz, M. (2008). Türkçede Okuduğunu Anlama Becerilerini Gelistirme Yolları. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, Sayı:9, s. 131 - 139.
- Zeybekoglu, S., Günes, A. B., & Yalçın, E. (2023). Examination of the Differential Item Functioning in the PISA 2018 Reading Test Items *Education and Science*. 1-23. <http://dx.doi.org/10.15390/EB.2023.12123>
- Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. D., & Gelin, M. N. (2005). A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological/Community Moderated (or Mediated) Test and Item Bias. *Journal of Educational Research & Policy Studies*, 5(1), 1-23.