Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 14(1) (2022) 1004–1028

IJCI
International Journal of Curriculum and Instruction

Reliability analysis of PISA 2018 reading literacy student questionnaire based on Item Response Theory (IRT): Turkey sample

Murat Polat [a] *, Çetin Toraman[b], Nihan Sölpük Turhan[c]

[a] *Anadolu University SFL, Tepebasi, Eskisehir, 26400, Turkey*
[b] *ÇOMÜMedical Faculty, Terzioglu Campus, Çanakkale,18200, Turkey*
[b] *FSMU Education Faculty, Üsküdar, Istanbul, 34840, Turkey*

Abstract

PISA (Program for International Student Assessment) tests have enabled the OECD countries to see not only the success of their students in gaining the ability to solve some daily problems they may encounter in their lives but also the place in the world rankings as a result of an objective evaluation comparing the achievement results of participant countries. Therefore, it is very important for the reliability and prestige of the program that the PISA exams and related student survey results are as unbiassed and error-free as possible. In this paper, validity and reliability study of the PISA 2018 Reading Literacy Student Survey (RLSS) Turkey sample was conducted including the analysis of survey items within the framework of Item Response Theory (IRT). The item fit indexes, item parameters according to the GPCM (Generalized Partial Credit Model), standard error values and item characteristics' reliability of the survey were determined via IRT, respectively. The data set included the answers of 6111 15-year-old Turkish students participated in PISA 2018. In the data analysis, through using the local independence assumption, Q3 statistics; IRT calibrations were tested with the help of the "Mirt v.1.30" program within the scope of "R v.4.0.5". In the study, each set of questions in the PISA student survey was examined independently from each other and each question set was considered as a separate attitude scale. The results showed that, although some of the items in PISA 2018 (RLSS) gave low level information, all the question sets in the test provided an acceptable model fit according to the GPCM. Upon examining the item characteristic curves, it was understood that the survey items showed valid and reliable results for testing different ability levels.

Keywords: PISA; IRT; student survey; reading literacy; reliability

* Murat Polat. ORCID ID.: https://orcid.org/0000-0001-5851-2322
E-mail address: mpolat@anadolu.edu.tr

1. Introduction

1.1. Background to the problem

Test results obtained from PISA and TIMSS (Trends in International Mathematics and Science Study) exams are driven from skill-oriented multiple-choice exams containing objective data. They are financed and administered by the Organization of Economic Cooperation and Development-OECD, of which Turkey is also a stakeholder, can be used to make comparisons related to educational success internationally (Woessmann, 2016). Of these two world-wide popular tests, the entire organization related to PISA exam is affiliated with the OECD, and this organization announced that the PISA exam is the most comprehensive education scale in this field with an international popularity and wide participation (OECD, 2004). Therefore, PISA exam results are seen as one of the prominent high-stakes tests that have a significant impact on administrators making critical decisions on world education policies (Bulle, 2011). In addition, PISA results have not only been used in analyses related to the measurement of education, but also paved the way for focusing more on different fields, such as educational economics, policies and philosophy (Waldow, 2013).

The PISA program started in 2000 and Turkey joined in this organization in 2003. The general aim of the program is to reveal the level of problem-solving skills of 15-year-old students (in the light of the knowledge and skills they gained from the education they receive at schools) through tests containing different academic skills (Özdemir, 2016). For this reason, in each PISA exam; one of the disciplines among reading, mathematics and science is chosen as the basis for the tests at the knowledge and skill level, and of these subjects one single subject is given more importance than others (OECD, 2004, 2007, 2010). For example, reading was the basic subject in the PISA application made in 2000, while a different subject (mainly mathematics was determined) was chosen in 2003. Then, in 2006, science became the core issue, while in 2009, reading skills were emphasized again. In this way, in the PISA tests that are carried out every three years, large-scale and highly valid measurement models were used by focusing on different subject skills (Taş et al., 2016). A total of 79 countries are currently participating in this program, and PISA test is administered once every three years in all these participating countries. What is more, contrary to what is believed, PISA tests are not limited to the content of the courses that students take at schools, but reveal students' cognitive skills in a specific subject. Thanks to the skill-focused student surveys which the participants answer after displaying their knowledge and abilities in the PISA tests by questioning both their test-focused and skill-focused attitudes in detail (OECD, 2008, 2013).

Critical findings and comparisons regarding the education quality of the countries participating in the program have emerged from the conducted studies using the results of the PISA exam, which has gained a worldwide participation and prestige since its first

implementation (Özer, 2020; Yalçın & Tavşancıl, 2014). Student questionnaires on average scores in PISA, country rankings, their status compared to the OECD average, and the test-related issues (distribution of the sample and skills included in the test) are used to assess and compare the literacy skills of participating students, and these analyzes provide different perspectives on the abilities, readiness and preparedness of pupils to real-life situations (Woessmann, 2016).

In Turkish context, test results obtained since 2003 have been the subject of many different studies. The differences in education quality and skill levels at both secondary and high schools have been considered as an important independent variable (Ataş & Karadağ, 2017; Berberoğlu & Kalender, 2005; Farmer, 2006). Another main finding of these studies is that the remarkable differences in the achievement scores of the schools included in the sample according to Turkey's PISA results is due to the heterogeneity of student profile diversity stemmed from the variety in Turkish school types and the entrance scores demanded during school enrollment, which differ significantly according to the school type. (Akın, 2015; Albayrak, 2009; Ataş & Karadağ, 2017; Berberoğlu & Kalender, 2005; Çiftçi, 2006; Erdoğan, 2018; Farmer, 2006). In addition to the research reports, student surveys' data obtained after each new PISA test is shared with academicians on-line free of charge. The skill-oriented student surveys are analyzed by program analysts with quantitative techniques, to reveal significant differences and/or similarities when compared with other countries about the literacy and focus skill levels of the participant students, by which they reveal their cognitive abilities. These opportunities shows how vivid and accountable the PISA program is to its stakeholders in terms of its distinguished merits such as objective measurement-evaluation and transparency in reporting (Özdemir, 2016).

While analyzing a single a country or drawing comparisons between various countries, the participant students in the survey are classified into ethnic origin, culture, language, etc. divisions. By the help of this classification, significant predictions can be obtained, provided that some adaptations are made on the items in terms of test-taker characteristics (Van de Vijver & Tanzer, 2004). In addition, studies carried out by analyzing the data of PISA student questionnaires are of great importance in terms of measuring the academic achievements of the pupils and lead to further discussion on education management, education planning and education philosophy (Waldow, 2013).

1..2. 2018 PISA Student Survey Turkey Sample

The PISA 2018 data of Turkey analyzed by researchers considering the achievement results of the students in different subjects, the status of the school where the students study and the attitudes of the students towards the education they receive, mostly lagged

behind the national success debates. However, the relationship of many variables such as school climate and students' well-being, life-discipline, life-satisfaction, bullying at school, teaching practices, should be discussed along with student success which has also been consistently revealed in the literature. Therefore, a careful examination of the findings on these variables could provide important information about the education system and the students' school life and climate.

Next, talking of the research side of PISA exams, there are a number of statistical methods, sampling and testing techniques be emphasized talking of PISA student surveys. It is common to observe certain sampling techniques used to determine the sample group, and the research results revealed according to the findings driven from this sampling (Özdemir, 2016). To illustrate, in some PISA student survey analyses, the sample weight has been considered as one of the most important factors affecting the results of the research (Grek, 2012). The main criterion determined by the PISA program while forming the sample groups for the tests is the sampling selection applied by weighting the size, location, type and economic status of the school apart from the other schools which have 15-year-old students at different classes in the participating countries. However, this weighting varies depending on the country, and standard criteria that can be applied to every country have not been determined yet (OECD, 2009). Liou and Hung (2015) criticized this situation in their studies and stated that they had doubts about the reliability of the PISA tests, especially on sampling techniques of PISA and TIMSS tests. Meanwhile, for Turkey, two main criteria have been determined for sampling and school selection. When these criteria are examined, the school type and socioeconomic region the school is in are considered as independent variables that the PISA program focuses on (OECD, 2009). To draw the study sample, different school types and one single school from each region are included in the sample set, and it is assumed that 35 participants who voluntarily took the PISA test among the 15-year-old students in a school represent the universe in that school (OECD, 2009). Yet, it has been stated in some studies that there is a need for in-depth research concerning the extent to which this sampling method represents the relevant universe (Strand & Demie, 2007; Freitas et al., 2015).

Another important point to be considered in the analyzes made on the PISA tests is the relevant data issue (Yalçın & Tavşancıl, 2014). As for the PISA student survey results, it is possible to access large data sets including all the achievement tests and student surveys via the link on the OECD website. Each set belongs to a different sample group consisting of participants which are assumed to represent the universe. Another feature of the PISA exam is that in the two-hour-exams to test mathematics, science and reading, not are all students tested on the same question booklets but does each student take an individual test on the common subject and the results of these tests are strengthened by student survey findings (OECD, 2009; Rutkowski et al., 2010).

In the literature, particularly in the last 10 years, it has been observed that in PISA data analysis, IRT is mostly used for validity and reliability studies. The reason for this may be the type of analyzes to document the item parameters which are determined independently of the respondent group. Similarly, the group characteristics are also determined independently of the item sample (Embretson & Reise, 2000). In addition, thanks to IRT, test results for each respondent can be examined and standard error estimations can be made separately. Doing this, even if respondents are tested with different questions, a standard framework for ability estimations according to IRT can be revealed (Hambleton; Swaminathan & Rogers, 1991). Therefore, test statistics (individuals' test scores, item average, item discrimination power, test validity and reliability, etc.) can be analyzed without being dependent on the group to which the test is applied, and without being dependent on the items of the test applied to each individual (Nartgün, 2002).

On the other hand, although the IRT is a more advanced application than the CTT (Classical Test Theory), it is not used widely by test-designers due to the difficulties in its application and interpretation. Today, in order to eliminate the problems experienced in test development and the necessary pilot studies based on the IRT, different and effective applications such as forming a large item pool, determining item bias, weighting the options and equalizing the test can help address the difficulties experienced in preparing large-scale multiple-choice tests compared to the ones barely developed in 2000s (Hambelton and Swaminathan, 1985).

In addition, it should also be reminded that there are two basic assumptions in IRT based measurement model, unidimensionality and local independence (De Ayala, 2009). While a unidimensional model assumes that the items in the test measure only a single ability, local independence on the other hand assumes that the items are independent from each other at the same ability levels (DeMars, 2010; Hambleton & Swaminathan, 1985). From this point of view, it is very important to examine and reveal the research results of each different PISA question set in terms of unidimensionality and local independence considering the purpose, item bias and item difficulty (Stout, 1990).

To summarize, the literature review documented for the present study revealed that no scientific studies were conducted on item discrimination, validity and reliability of the of the PISA 2018 RLSS Turkey student survey based on item response theory (IRT). For this reason, in this study, PISA 2018 RLSS student survey's validity and reliability analyses were conducted on Turkey data according to the IRT. Considering this objective, reliability analyzes of the PISA 2018 RLSS items are valuable for researchers for it can shed light on other surveys on the relevant data set and contribute to scientific predictions in determining student attitudes towards the PISA tests. It was also aimed to provide in-depth analyzes of the mentioned survey made for other PISA researchers, who consider those tests as the supporting data tools to strengthen the findings of PISA tests.

2. Method

In this study, which was carried out by adopting the screening model, analyzes on item discrimination, validity and reliability were made on the Turkish sample of the PISA 2018 RLSS. The screening model helps researchers who aim to portray a past or present phenomenon as it actually is, without affecting the variables or trying to influence them (Karasar, 2014). The situation that is the subject of the research is aimed to be shown in its own context, without any interventions.

2.1. Instruments

PISA 2018 (RLSS) Turkey data set, the subject of the research, was a multiple-choice measurement tool, downloaded from the relevant website of the OECD, and consisted of five grades and three sets of questions. There were 6 items in the first set (ST164), 5 items in the second set (ST165) and 5 items in the third set of questions (ST166).

2.2. Participants

Participants of the PISA 2018 RLSS Turkey data were students aged 15 and over in grades 7 or above, studying in different types of schools around Turkey. A total of 6890 students participated in this survey. 779 participants' data were excluded (as it was done in similar cases in the literature) from the study because of the missing values in the data set (Kline, 2005). After this process, a total of 6111 participants' test data was analyzed. Of the students in the data set, 18 (0.3%) were 8th grade, 1131 (18.5%) were 9th grade, 4775 (78.1%) were 10th grade, 183 (3%) were 11th grade and 4 (0.1%) were 12th graders. The data used in the study was the original dataset downloaded from the OECD PISA web-site. No stratification or grouping was made in Turkey dataset by the researchers.

2.3. Data Analysis

In the first stage, the missing data was all removed and the data of 6111 participants in total remained in the data set. In the second stage, three sets of survey questions containing the answers to the reading survey were considered as a single measurement tool and subjected to IRT-based factor analysis. Next, each set of survey questions were assumed as independent scales. Thus, ST164, ST165 and ST166 question sets were considered and analyzed independent of each other.

For Likert-type scale items in the response set, it was suggested to examine the assumptions of unidimensionality and local independence in validity and reliability analyses with IRT (Rosseel, Y. (2012; Zhao, 2008). Therefore, unidimensionality was analyzed with item correlation matrix, exploratory and confirmatory factor analysis

(Hambleton, Swaminathan & Rogers, 1991). In addition, the local independence assumption was tested using the Q3 statistics (Yen, 1993), the IRT calibrations were made with the "Mirt v.1.30" (Chalmers, 2012) package program in the R v.4.0.5 Software, and the findings were presented respectively.

3. Results

3.1. Unidimensionality Test

There was a total of 16 items measuring students' reading skills in three sets of questions in the PISA 2018 *CRSS*. Certain assumptions need to be examined in order to be able to analyze the survey items based on the IRT model (Cokluk et al., 2016; Erdogan & Guvendir, 2019; Gelbal, 1994; Teacher, 1995; Rajchert et al., 2014; Shala & Grajcevci, 2018). First, the correlation matrix of the survey items was created and then the unidimensionality assumption was tested. The correlation matrix findings were shown in Table 1.

Table 1. Reading Skills Student Survey Items' Correlation Matrix

C		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	ST164Q01IA	1														
2	ST164Q02IA	0,341	1													
3	ST164Q03IA	0,182	0,217	1												
4	ST164Q04IA	0,199	0,181	0,413	1											
5	ST164Q05IA	0,188	0,180	0,405	0,494	1										
6	ST164Q06IA	0,082	0,174	0,377	0,313	0,363	1									
7	ST165Q01IA	0,250	0,228	0,273	0,311	0,263	0,221	1								
8	ST165Q02IA	0,247	0,258	0,147	0,173	0,126	0,185	0,446	1							
9	ST165Q03IA	0,224	0,267	0,264	0,352	0,262	0,247	0,384	0,352	1						
10	ST165Q04IA	0,195	0,179	0,331	0,418	0,404	0,243	0,394	0,246	0,494	1					
11	ST165Q05IA	0,162	0,140	0,312	0,516	0,426	0,245	0,338	0,187	0,419	0,625	1				
12	ST166Q01HA	0,163	0,154	0,120	0,085	0,091	0,095	0,163	0,184	0,112	0,085	0,070	1			
13	ST166Q02HA	0,152	0,135	0,236	0,219	0,232	0,110	0,210	0,093	0,218	0,321	0,299	0,417	1		
14	ST166Q03HA	0,172	0,173	0,090	0,030	0,034	0,118	0,133	0,241	0,117	0,004	-0,005	0,501	0,195	1	
15	ST166Q04HA	0,092	0,092	0,080	0,048	0,085	0,091	0,095	0,076	0,110	0,113	0,102	0,011	0,223	0,104	1
16	ST166Q05HA	0,111	0,122	0,222	0,208	0,220	0,140	0,186	0,082	0,210	0,312	0,283	0,272	0,581	0,152	0,258

N=6111

When the correlations between the items in the matrix were examined, it was determined that 16 items of the PISA 2018 CRSS did not gain correlation values like a single measurement tool nor revealed high correlation levels. For this reason, the items of the survey related to reading skills could not meet the unidimensionality assumption. Therefore, in order to make more effective comparisons IRT-based analyzes, in which the three sets of survey questions were considered as separate constructs, were conducted.

3.2. The *model in which 3 survey sets were considered as separate measurement tools*

After it was understood that the model was not unidimensional (it was the first assumption for IRT and was checked) the other assumption (local independence) test was initiated on student survey's 3 sets. Then, with the use of Q3 test on PISA 2018 CRSS, it was revealed that there was no item which impairs local independence among the 16 items. Based on this finding, item calibrations were determined with the Generalized Partial Credit Model (GPCM) within the scope of IRT for the survey items. S_χ2, (degree of freedom), RMSEA and level of significance statistics of the items according to GPCM were carried out (Cheung & Rensvold, 2002). The results were presented in Table 2.

Table 2. PISA 2018 CRSS Item Fit Index Based on IRT

Items	**GPCM**		
	S_χ^2	**Df**	**RMSEA**
Item 1	491.383	299	0.010
Item 2	518.870	293	0.011
Item 3	394.597	263	0.009
Item 4	449.973	240	0.012
Item 5	343.484	254	0.008
Item 6	376.974	281	0.007
Item 7	369.754	259	0.008
Item 8	552.957	282	0.013
Item 9	416.140	251	0.010
Item 10	417.672	221	0.012
Item 11	425.267	220	0.012
Item 12	565.329	303	0.012
Item 13	503.777	282	0.011
Item 14	814.885	304	0.017
Item 15	443.396	307	0.009
Item 16	416.783	285	0.009

The critical level for the RMSEA value, which is one of the most important fit indices for measurements made with IRT, is 0.08 and below; thus, this value indicated a good item fit for the model (Büyüköztürk et al., 2016; Stout, 1990). According to the item concordance statistics in Table 2, the RMSEA values for all the items were less than 0.08. Based on this result, it was decided that 16 items of the reading skills measurement tool provided a model fit according to the GPCM. In the next step, the "a" (item discrimination) and "b" (item difficulty) parameters and standard errors (standard errors) of the items whose model fit was determined according to the GPCM were studied separately for each item and the results were presented in Table 3.

Table 3. Item Parameters and Standard Error Values According to GPCM

Item	**a(SE)**	**b1(SE)**	**b2(SE)**	**b3(SE)**	**b4(SE)**	**b5(SE)**
Item 1	0.252(0.011)	-1.582(0.184)	-0.494(0.163)	0.943(0.175)	1.567(0.204)	-1.536(0.215)
Item 2	0.268(0.012)	-1.378(0.155)	0.571(0.152)	1.022(0.170)	0.996(0.186)	0.366(0.193)
Item 3	0.490(0.017)	-1.710(0.103)	-0.328(0.086)	0.192(0.088)	0.752(0.096)	0.041(0.099)
Item 4	0.689(0.023)	-1.644(0.104	-1.276(0.086)	-0.548(0.074)	-0.274(0.069)	-0.973(0.073)
Item 5	0.571(0.019)	-1.773(0.114)	-0.940(0.094)	-0.548(0.086)	-0.322(0.079)	-0.579(0.076)
Item 6	0.324(0.012)	-0.400(0.127)	0.346(0.135)	0.660(0.147)	0.745(0.159)	-0.878(0.162)
Item 7	0.521(0.017)	-1.059(0.085)	-0.178(0.079)	0.644(0.088)	0.820(0.099)	-0.103(0.102)
Item 8	0.322(0.013)	-0.588(0.125)	-0.031(0.124)	0.953(0.138)	1.607(0.165)	0.318(0.173)
Item 9	0.645(0.021)	-2.053(0.099)	-0.907(0.072)	-0.177(0.066)	0.443(0.068)	0.103(0.070)
Item 10	0.993(0.033)	-1.808(0.077)	-1.137(0.060)	-0.661(0.051)	-0.138(0.045)	-0.104(0.045)
Item 11	0.906(0.031)	-2.026(0.088)	-1.039(0.067)	-0.555(0.061)	-0.432(0.056)	-0.819(0.059)
Item 12	0.167(0.009)	3.375(0.310)	0.398(0.269)	2.067(0.316)	1.847(0.355)	-3.305(0.391)
Item 13	0.333(0.013)	-0.888(0.147)	-0.258(0.144)	0.061(0.147)	0.311(0.151)	-2.328(0.169)
Item 14	0.139(0.009)	6.211(0.514)	0.424(0.322)	2.988(0.409)	3.981(0.517)	-1.736(0.509
Item 15	0.123(0.009)	3.584(0.398)	2.657(0.398)	4.141(0.516)	3.266(0.571)	-5.521(0.665)
Item 16	0.298(0.012)	0.042(0.169)	-0.517(0.169)	0.152(0.170)	0.004(0.171)	-2.811(0.193)
		Iteration=49	LogLikelihood= - 157542.192		p<.05	

In IRT, the distinctiveness value of an ideal scale item (ie the "a" parameter) should be between 0.5 and 2. In the literature, it is accepted that this parameter be between 0.75 and 2.50 (Flannery, Reise & Widaman, 1995). Table 3 values showed that the discrimination values of items 10 and 11 were at the desired levels. The ideal (medium difficulty level) limits for item difficulty levels (that is, the "b" parameter) were considered to be between -1.00 and 1.00 (Hambleton, 1994). In ability or achievement tests, items with less than a -1.00-difficulty level are considered easy, and items over 1.00 are considered difficult. Items 6 and 8 gained value within the desired item difficulty

parameters. The analyis made according to the GPCM (LogLikelihood, p<.05) proven the consistency of the survey items. Next, item characteristic curves of the items included in the data set were shown in Figure 1.

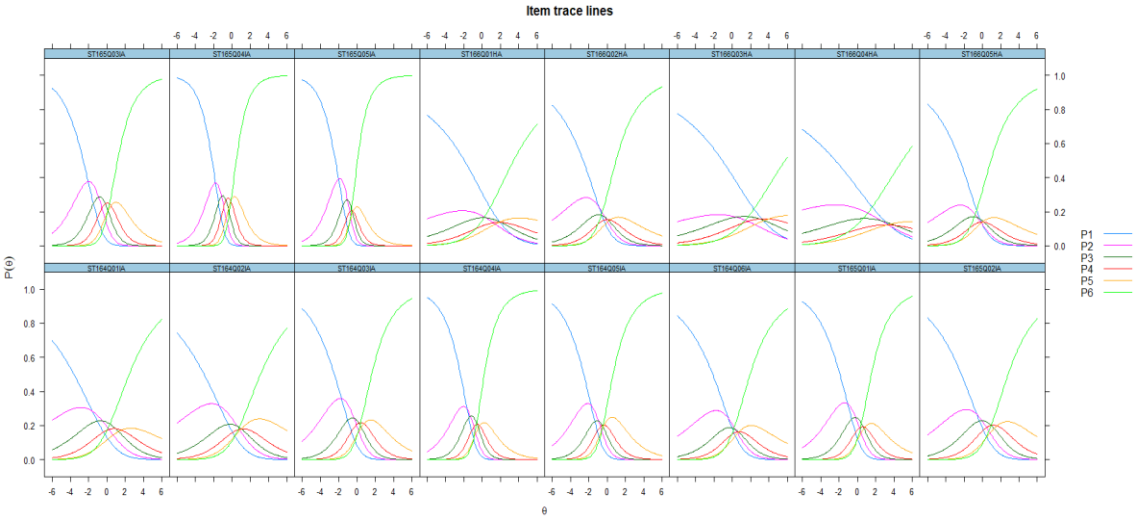

Figure 1. Item characteristic curves of the PISA 2018 CRSS

According to the item characteristic curves in Figure 1, it was seen that the items in the survey, together with the distractors, worked well and were highly distinctive in different cognitive levels for the target subject (reading skills). The discrimination level of the response categories of the item ST164Q01IA (Item 1), ST164Q02IA (Item 2), ST166Q01HA (Item 12), ST166Q03HA (Item 14), ST166Q04HA (Item 15) was relatively lower than the remaining items. In the light of these findings, it can be said that the response categories of the items in the survey were recognized by the participants and the items had distinctive qualities. Next, item information functions were shown in Figure 2.

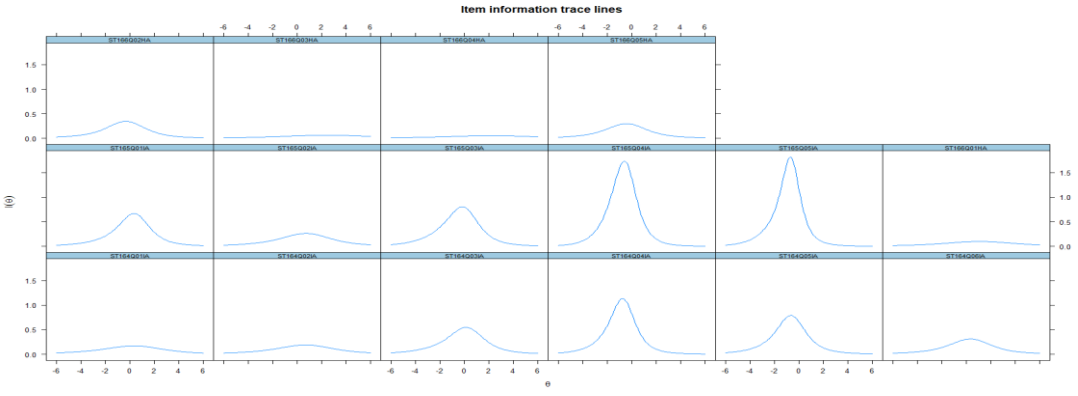

Figure 2. Item information functions of the PISA 2018 CRSS

The item information function is a graphical representation showing the range of features (the feature that is tried to be measured in the scale) by which the item best distinguishes students who take the test (Edelen & Reeve, 2007). In the item information function, the higher the peak of the curve, the more information the item gives. When the item information functions of the PISA student survey items were examined, the least functional or informative items were found as ST164Q01IA, ST164Q02IA, ST164Q03IA, ST164Q06IA, ST165Q02IA, ST166Q01HA, ST166Q02HA, ST166Q03HA, ST166Q04HA and ST166Q05HA. Next, the test information function was shown in Figure 3.

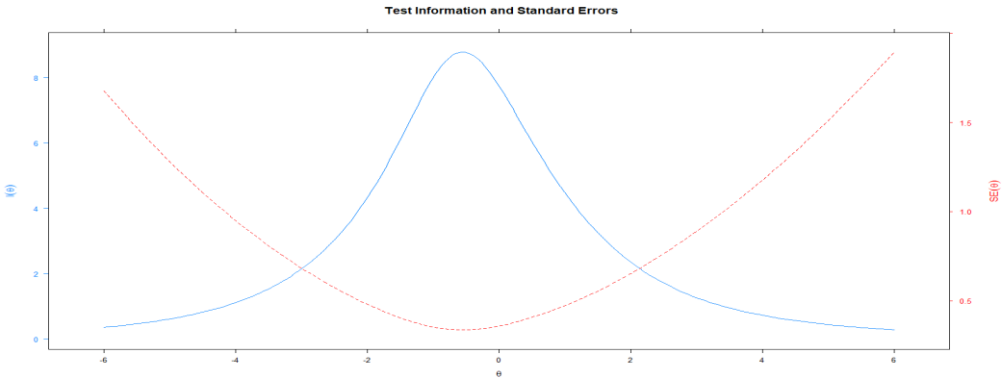

Figure 3. Information functions of the PISA 2018 CRSS items

Information function measurement tool shows the level of a single item's giving information about the feature it tests (Hambleton, Swaminathan & Rogers, 1991). As could be seen in Figure 3, the appearance of a normal distribution curve as a result of the analysis is an indication that the measurement tool gives information on different levels of the measured feature. The level that the measurement tool gives the best information was between -2 and 2 (Tabachnick & Fidell, 1996; Wood, 2008). In addition, the marginal reliability coefficient of this measurement tool was also calculated as 0.85, which indicated a high reliability level.

3.3. Validity and Reliability Results of ST164 Item Cluster According to IRT

According to the Q3 test, it was found that among the 6 items in the ST164 item set in the PISA 2018 CRSS, there was no item that impairs local independence. For the items of the PISA 2018 student survey, ST164 item test's item calibrations were determined with the Generalized Partial Credit Model (GPCM) within the scope of IRT. Then, S_χ2, (degree of freedom), RMSEA and level of significance statistics of the items according to GPCM were made. The results were shown in Table 4.

Table 4. Item fit indexes according to IRT for the first set of questions (ST164)

Item Set ST164	**GPCM**		
	S_χ^2	**df**	**RMSEA**
ST164Q01IA	403.472	98	0.023
ST164Q02IA	498.192	96	0.026
ST164Q03IA	279.071	90	0.019
ST164Q04IA	335.872	86	0.022
ST164Q05IA	266.065	86	0.019
ST164Q06IA	326.915	93	0.020

According to the item concordance statistics in Table 4, the RMSEA values of all the items in the cluster were less than 0.08. Based on this result, it was decided that the ST164 question set provided model fit according to GPCM. In the next step, the "a" and "b" parameters and standard errors of the items whose model fit was analyzed according to GPCM were estimated separately for each item and results were presented in Table 5.

Table 5. Item parameters and standard error values according to GPCM for the first set (ST164)

Item Set ST164	**a(SE)**	**b1(SE)**	**b2(SE)**	**b3(SE)**	**b4(SE)**	**b5(SE)**
ST164Q01IA	0.211(0.011)	-1.769(0.222)	-0.539(0.194	1.116(0.211)	1.801(0.247)	-1.962(0.266)
ST164Q02IA	0.241(0.012)	-1.480(0.174)	0.652(0.171)	1.122(0.191)	1.059(0.208)	0.311(0.215)
ST164Q03IA	0.677(0.026)	-1.554(0.077)	-0.400(0.063)	0.117(0.065)	0.669(0.071)	0.312(0.075)
ST164Q04IA	0.821(0.035)	-1.644(0.088)	-1.268(0.074)	-0.585(0.063)	-0.267(0.059)	-0.738(0.068)
ST164Q05IA	0.849(0.036)	-1.680(0.079)	-0.951(0.065)	-0.537(0.059)	-0.222(0.055)	-0.193(0.056)
ST164Q06IA	0.426(0.017)	-0.508(0.098)	0.169(0.103)	0.514(0.113)	0.687(0.122)	-0.438(0.126)
Iteration=14		LogLikelihood= - 60442.634		p<.05		

Findings in Table 5 revealed that the discrimination values of ST164Q04IA and ST164Q05IA were at the desired levels. It was observed that the item difficulty parameters of the ST164Q06IA were in the range of the sought-after degrees. Moreover, the estimations made according to the GPCM (LogLikelihood, $p<.05$) proven the consistency of the measurement tools. Next, the item characteristic curves of the items in the question set were given in Figure 4.

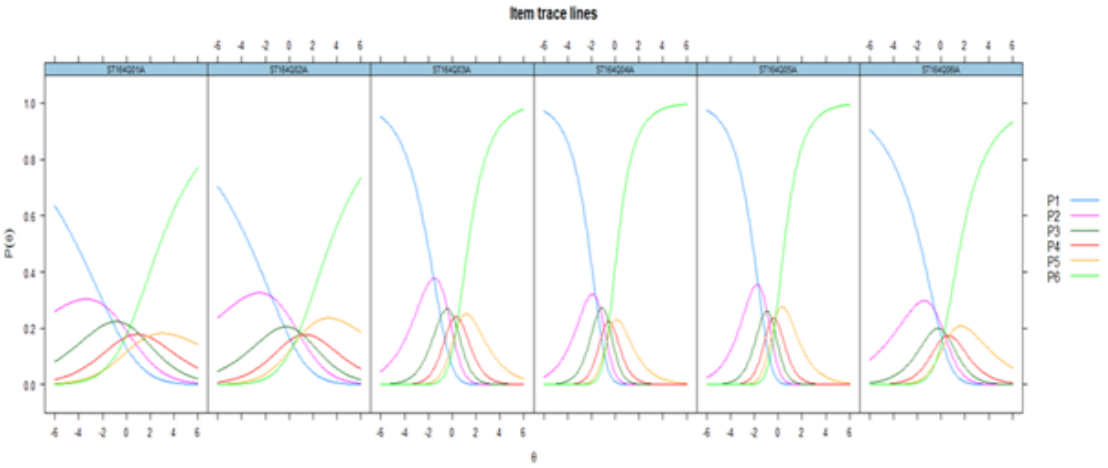

Figure 4. Item characteristic curves of the first question group (ST164)

According to the item characteristic curves seen in Figure 4, it was noticed that items in set ST164, together with their distractors, worked well on different cognitive levels and were distinctive. The discrimination of the response categories of item ST164Q01IA and ST164Q02IA was relatively lower than the remaining items. Response categories of the items in the measurement tool were understood by the participants and they served as distinguishing features. Item information functions were shown in Figure 5.

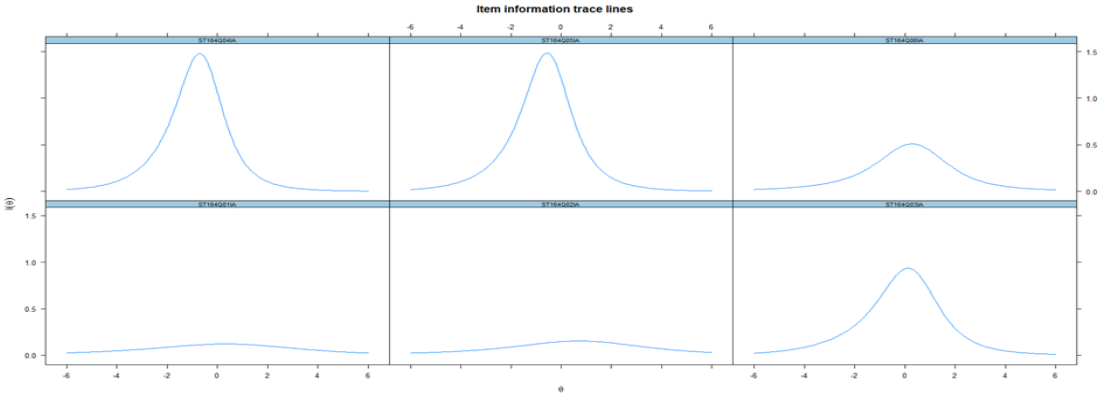

Figure 5. Item information functions of the first question group (ST164)

When the item information functions of the ST164 question items were examined, the least informative items were found as ST164Q01IA, ST164Q02IA, ST164Q03IA, ST164Q06IA, respectively. The test information function was shown in Figure 6.

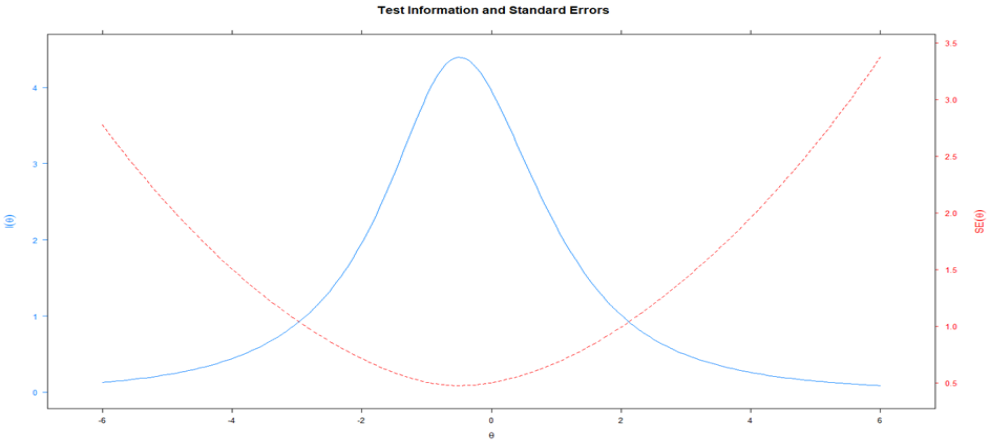

Figure 6. Information functions of the first question group (ST164)

According to the obtained analysis results, the ST164 question set was a good measurement tool that gave sufficient information about the feature it aimed to measure. The level that the measurement tool gives the best information was between -2 and 2. In addition, the marginal reliability coefficient of this measurement tool was calculated as 0.74. Next, the fit indices calculated according to GPCM for the ST165 question set in the next analysis were shown in Table 6.

Table 6. Item fit indexes according to IRT for the second set (ST165)

ST165 Item Cluster	**GPCM**		
	S_χ^2	**df**	**RMSEA**
ST165Q01IA	446.677	68	0.030
ST165Q02IA	623.617	71	0.036
ST165Q03IA	454.860	69	0.030
ST165Q04IA	350.550	63	0.027
ST165Q05IA	356.751	65	0.027

According to the item concordance statistics in Table 6, the RMSEA values were less than the critical value of 0.08. According to this result, it was decided that the ST165 question set provided model fit according to GPCM. Based on the results obtained, the "a" and "b" parameters and standard errors of the items whose model fit was determined according to GPCM were estimated and the analysis results were presented in Table 7.

Table 7. Item parameters and standard error values according to GPCM for set ST165

ST165 Item Cluster	a(SE)	b1(SE)	b2(SE)	b3(SE)	b4(SE)	b5(SE)
ST165Q01IA	0.531(0.020)	-1.083(0.083)	-0.198(0.078)	0.633(0.087)	0.831(0.098)	-0.066(0.102)
ST165Q02IA	0.349(0.015)	-0.614(0.116)	-0.056(0.115)	0.895(0.128)	1.536(0.153)	0.378(0.160)
ST165Q03IA	0.755(0.027)	-1.980(0.086)	-0.902(0.062)	-0.195(0.057)	0.421(0.059)	0.224(0.061)
ST165Q04IA	1.393(0.061)	-1.777(0.059)	-1.118(0.046)	-0.641(0.038)	-0.129(0.034)	0.090(0.036)
ST165Q05IA	0.938(0.037)	-2.037(0.085)	-1.076(0.065)	-0.586(0.060)	-0.438(0.055)	-0.771(0.061)
	Iteration=36	LogLikelihood= - 48016.368		p<.05		

Statistical analysis which were made according to GPCM (LogLikelihood, p<.05) proven the consistency of the items in the item set ST165. Item characteristic curves were shown in Figure 7.

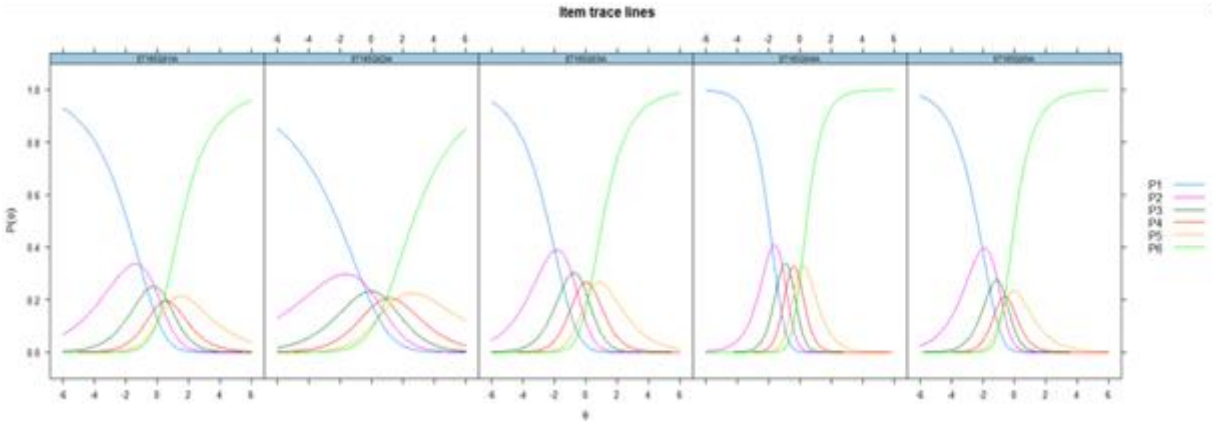

Figure 7. Item characteristic curves of the 2nd Item Set (ST165)

According to the item characteristic curves in Figure 7, it was seen that the items in the ST165 item set, together with their distractors, functioned well and were distinctive for different levels of cognitive reading skills. The discrimination of the response categories of item ST165Q02IA was relatively lower than the remaining items. The results of the analysis proven that the response categories of the items in the measurement tool were well understood by the participants and had a distinctive function. The item information functions of the ST165 question cluster were shown in Figure 8.

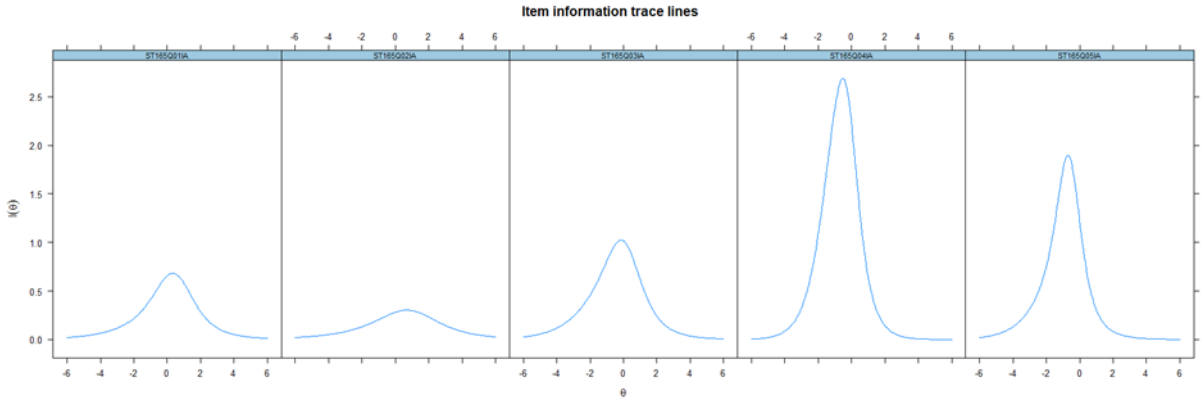

Figure 8. Item information functions of the 2nd Item Set (ST165)

When the item information functions of ST165 were examined, it can be seen that the least informative items were ST165Q01IA, ST165Q02IA and ST165Q03IA. Next, the test information function related to the ST165 question set was shown in Figure 9.

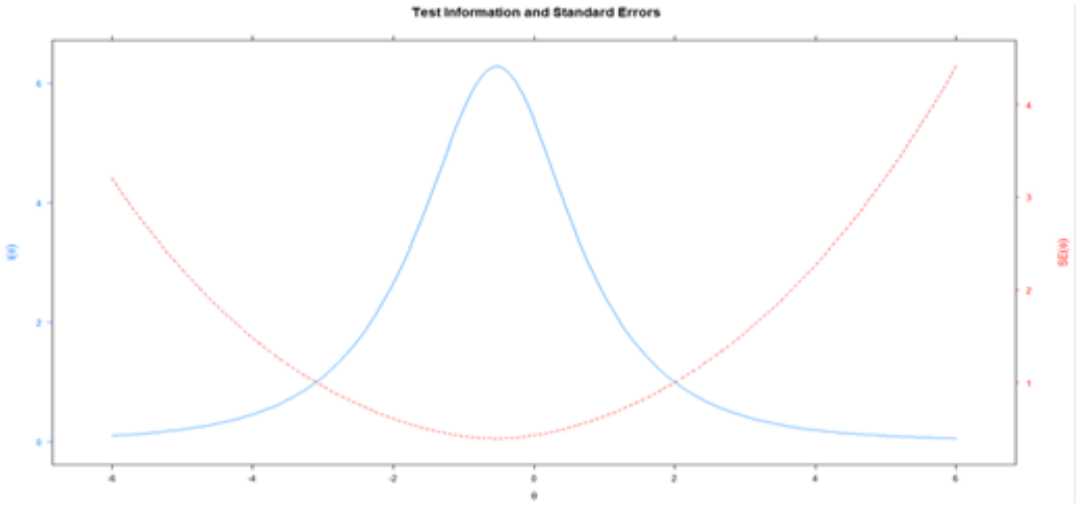

Figure 9. Information functions of the 2nd item set (ST165)

According to the analysis results, the ST165 question set was a good measurement tool that gave sufficient information about the feature it aims to measure. The level that the measurement tool gave information was between -2 and 2. The marginal reliability coefficient for the measurement tool was calculated as 0.81. Finally, the fit indices calculated according to GPCM for the ST166 question set of the PISA reading activity test were presented in Table 8.

Table 8. Item fit indexes (fit-indexes) according to IRT for the third set (ST166)

ST166 tem Cluster	**GPCM**		
	S_χ^2	**df**	**RMSEA**
ST166Q01HA	829.696	74	0.041
ST166Q02HA	619.035	66	0.037
ST166Q03HA	1212.415	74	0.050
ST166Q04HA	1009.134	74	0.045
ST166Q05HA	441.303	68	0.030

According to the item concordance statistics in Table 8, the RMSEA values of the items were less than 0.08. According to this result, it was decided that the ST166 question set provided model fit according to GPCM. The "a" and "b" parameters and standard errors of the items whose model fit was determined according to GPCM were analyzed and the results were shown in Table 9.

Table 9. Item parameters and standard error values according to GPCM for ST166

ST166 Item Cluster	**a(SE)**	**b1(SE)**	**b2(SE)**	**b3(SE)**	**b4(SE)**	**b5(SE)**
ST166Q01HA	0.425(0.017)	0.993(0.112)	0.064(0.107)	0.945(0.122)	1.075(0.138)	-0.729(0.147)
ST166Q02HA	1.666(0.118)	-1.107(0.037)	-0.520(0.034)	-0.132(0.034)	0.217(0.035)	0.084(0.050)
ST166Q03HA	0.252(0.013)	3.281(0.239)	0.236(0.178)	1.791(0.214)	2.478(0.270)	-0.536(0.278)
ST166Q04HA	0.182(0.010)	2.299(0.245)	1.759(0.256)	2.850(0.325)	2.347(0.373)	-3.474(0.414)
ST166Q05HA	0.693(0.027)	-0.611(0.076)	-0.568(0.075)	-0.039(0.075)	0.127(0.075)	-0.831(0.086)
Iteration=36		LogLikelihood= - 48953.471			p<.05	

The analysis which were made according to GPCM (LogLikelihood, p<.05) proven the consistency of the items in the measurement tool. Item characteristic curves were presented in Figure 10.

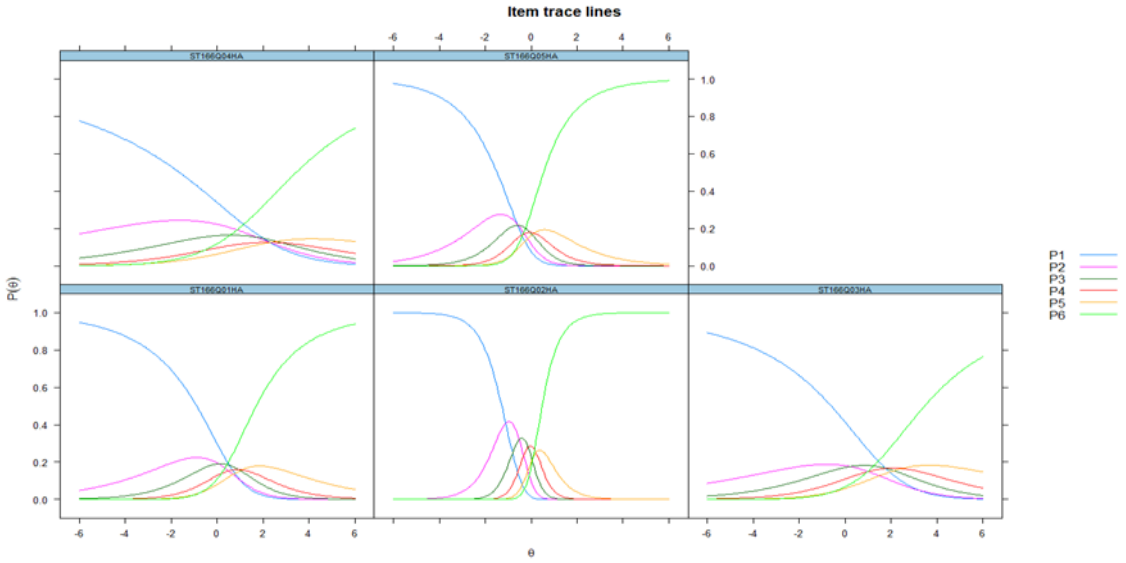

Figure 10. Item characteristic curves of the 3rd item set (ST166)

According to the item characteristic curves in Figure 10, it was seen that the items in the 3rd item cluster ST166, together with their distractors, functioned well and were distinctive to measure different levels of reading skills. The discrimination of the response categories of the item ST166Q03HA and ST166Q04HA was relatively lower than the other items. Response categories of the items in the measurement tool were recognized well by the participants and served as a distinguishing testing feature. Finally, item information functions were shown in Figure 11.

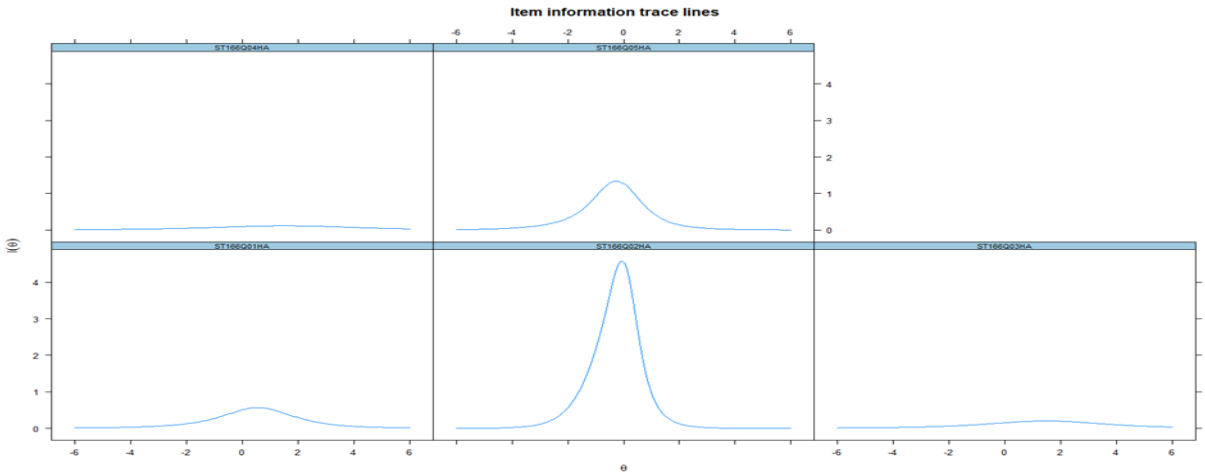

Figure 11. Item information functions of the 3rd item set (ST166)

When the item information functions of the item set ST166 were examined, it was observed that the least informative items were ST166Q01HA, ST166Q03HA, ST166Q04HA and ST166Q05HA compared to the remaining items in the same set. Set ST166 items test information function was shown in Figure 12.

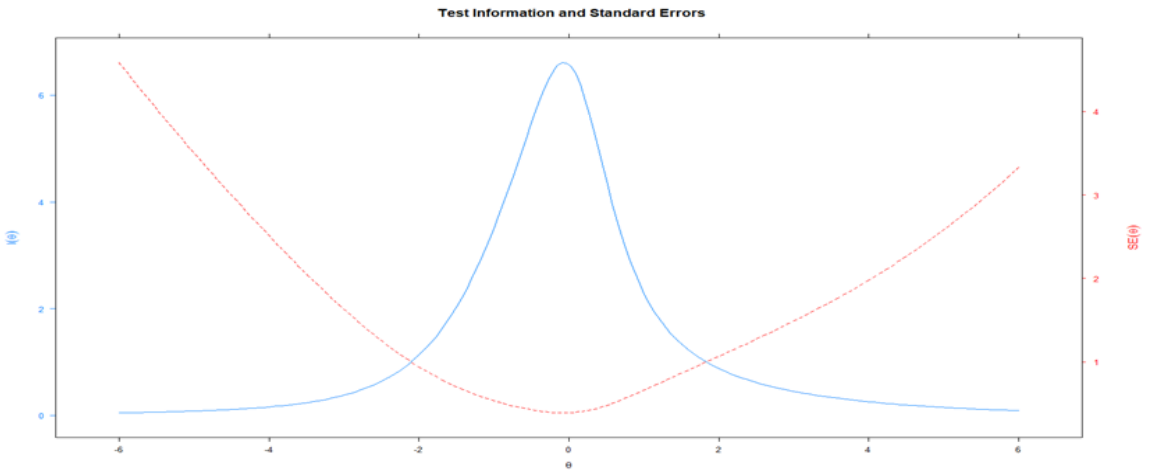

Figure 10. Information functions of the 3rd item set (ST166)

It was revealed that ST166 question set was a good measurement tool which gave sufficient information about the feature it was designed to test. The level that the measurement tool gives the best information was between -1.5 and 1.5. To conclude, the marginal reliability coefficient of the measurement tool was calculated as 0.76 which could be considered as an acceptable degree of.

4. Discussion & Conclusion

The aim of this research was to reveal the validity and reliability levels of the PISA 2018 Reading Literacy Student Survey (RLSS) according to the item analyzes made on Turkey data on the basis of IRT. With this aim in mind, the three sets of questions in the PISA 2018 student survey were structured independently of each other, and each question group; therefore, was accepted as an independent scale. In other words, ST164 question set with 6 items, ST165 question set with 5 items and finally ST166 question set with 5 items were considered as unique and independent scales. Camilli (2006) stated that investigating item fit indices in IRT studies, item parameters according to GPCM, standard error values, item characteristics and reliability degrees are effective methods for validity and reliability analysis of a test. Therefore, the above-mentioned analyses were applied respectively on each set of survey questions, and the results were presented separately in the findings section.

To begin with, while examining the validity and reliability of PISA 2018 Reading Literacy Student Survey (RLSS) items, question sets including ST164, ST165 and ST166 were analyzed according to IRT, and a number of important findings were driven. As a result of the item correlation matrix analysis on question sets ST164, ST165 and ST166, it was concluded that there were no unidimensional structure and according to Q3 test, there were no items that impaired local independence among the items in the survey. According to GPCM, it was determined that the items provided a successful model fit according to S_χ2, degrees of freedom, RMSEA and level of significance. In Ferhan's (2018) research, a similar finding was reported on the PISA 2012 data, especially through item analyzes made based on the IRT for mathematics scale, and it was concluded that the scale had a sound construct validity. Thus, the test information function values (which were estimated by considering different item weightings based on IRT) were determined, and it was concluded that using this method led us to drive more information on students' attitudes towards the PISA tests, strategies they adopted, and minimalization of possible measurements errors.

Next, it was observed that the discrimination of the response categories of ST164Q01IA and ST164Q02IA, which were among the items in the question set ST164 in the PISA 2018 RLSS, was relatively lower than the remaining items. Considering the discriminating features of the response categories of the item ST165Q02IA in question set ST165, it was determined that the discrimination of this item was relatively lower than the other set items. In addition, considering the discriminating features of the response categories of items ST166Q03HA and ST166Q04HA in question set ST166, the discrimination observed was found to be relatively lower than the remaining items in the same set. This result shows that a similar situation may have occurred for the items that were included in past PISA student surveys which had low discrimination statistics. As a result of the section evaluation made for the PISA 2015 RLSS, it was seen that the number of sets in the survey was increased from two to three by OECD (OECD, 2020). The reason for this change can be interpreted as the intent to upsurge the reliability level of the scale by increasing the number of items. Moreover, it was observed that the items used in the PISA 2018 RLSS were distinctive in terms of participants' answers. However, removing or changing the items with relatively low discrimination levels compared to the remaining items in the set could affect the discrimination features of the untouched items (Chen, 2007). For this reason, the item discrimination values obtained as a result of this study led us to conclude that pilot studies are essential for future PISA student surveys, because to what extent the answers given by the students to the survey questions (which are thought to reflect adnanced reading strategies) are as important as the students PISA test scores is still unknown since the results of PISA 2018 RLSS may still reflect the actual reading strategies participants use in real life. To illustrate, when different studies on this issue were examined, it was seen that students sometimes give wrong answers to the survey questions just because of the "supervisor effect" and prefer

to act in the way s/he was expected to act instead of reflecting researchers what actually happened in its PISA experience (Braten et al., 2004; Pintrich et al., 1992).

Finally, the item information function in the question sets ST164, ST165 and ST166 were examined. Initially, the items in the 1st question set (ST164) were examined in detail. Items ST164Q01IA, ST164Q02IA, ST164Q03IA and ST164Q06IA gave little information about the participant's reading skills. When the item information function of the question set ST165 was examined, it was seen that the least informative items were ST165Q01IA, ST165Q02IA and ST165Q03IA. The last but not the least, in set ST166, it was determined that the least informative items were ST166Q01HA, ST166Q03HA, ST166Q04HA and ST165Q05IA.

It was observed that test information function values, determined for different sets of questions according to the analyzes based on IRT, gave more information to students and measured their reading skills with less errors. Hopfenbeck and Maul (2011) shared a similar result in their study on PISA learning strategies survey. They reported that while students respond to the survey items in the PISA RLSS, participants, particularly who got low scores in the achievement test, focused mostly on strategies related to a specific reading component; therefore, they presented a similar poor performance in the achievement test. It is also stated that students' test performance is unpredictable while answering the PISA RLSS and this case could be an indicator of the fact that students have different literacy levels as they do in achievement tests, and this hypothesis may be tested in their achievement test scores using the survey results. On the other hand, if the students (participating in the PISA achievement tests) are not sufficiently informed about the importance of this survey which is generally administered after the achievement tests, the answers they give may not fully reflect their real performance they present in the achievement test, and in order to prevent such inconsistencies, conducting an interview instead of giving a questionnaire will give the students a chance to recall more about the test and reflect his/her opinions in depth.

As a result of this research it was found that the IRT models, designed to find out item discrimination and item difficulty of the survey questions in the PISA 2018 RLSS were satisfactory. It was also observed that the marginal reliability coefficients, which are the indicators of the reliability of the test, were also above acceptable levels (.70). Based on these results, it can be reported that the PISA RLSS was a valid and reliable measurement tool to investigate participants' reading strategies. However, it was also observed that the item information levels of 3 items in the first set, 2 items in the second, and 3 items in the third set of PISA RLSS were lower than the remaining items. Eventually, it can be recommended that if the PISA RLSS is to be used in the future, it is necessary to review the questions and their distractors with low distinctiveness levels, to let the experts make better revisions on relevant items, and give more emphasis on pilot studies of similar surveys. As a final word, it should be noted that, findings of this study

will be significant for the PISA program administration (called to provide more reliable data to the participating countries) to better compare OECD countries' education qualities, encourage researchers to compare the reasons of possible PISA test performance differences or similarities among countries, and see the cognitive profiles of the participating students more vividly.

References

Akın, C.A. (2015). Comparison of IRT-Likelihood Ratio, Ordinal Logistic Regression and Poly-Sibtest Methods in Determining Changing Item Function. *E-International Journal of Educational Research. 6(1)*, 1-16.

Albayrak, A. (2009). *Some factors affecting the science achievement of students in Turkey according to PISA 2006 exam results.* (Unp. Master's Thesis. Hacettepe Uni. Inst. of Soc. Sciences).

Ataş, D., & Karadağ, Ö. (2017). An analysis of Turkey's PISA 2015 results using two-level hierarchical linear modeling. *Journal of Language and Linguistic Studies, 13(2),* 720-727.

Berberoğlu, G., & Kalender, İ. (2005). Examining student achievement by years, school types, regions: ÖSS and PISA analysis. *Educational Sciences and Practice, 4(7),* 21-35.

Braten, I., Samuelstuen, M., & Strømsø, H. (2004). Do students' self-efficacy beliefs moderate the effects of performance on self-regulatory strategy use? *Educational Psychology, 24(2),* 231–247

Bulle, N. (2011) Comparing OECD educational models through the prism of PISA. *Comparative Education, 47 (4),* 503-521.

Büyüköztürk, Ş., Çokluk, Ö., and Köklü, N. (2016). *Statistics for the social sciences (*18th Edition). Ankara: Pegem Academy.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.221-256). Westport: American Council on Education and Praeger Publishers

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48(6*).

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14(3),* 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9(2),* 233–55.

Cakici, D.E. (2013). Analysis of Bi-Category Data Obtained from PISA 2009 Reading Test with Bılog Program. *Journal of Research in Education and Teaching. Volume:2/4.* ISSN: 2146-9199.

Farmer, A. (2006). *Examination of some factors affecting the success of students in Turkey according to the results of PISA 2003 exam mathematics subtest.* (Unpublished Master's thesis. Hacettepe University Institute of Social Sciences).

Çokluk, Ö., Şekercioğlu G. and Büyüköztürk, Ş. (2016). *Multivariate statistics, SPSS and lyrical applications for social sciences* (4th Edition). Ankara: Pegem Academy.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. Guilford Press. U.S.A.

DeMars, C. (2010). *Item response theory.* (H. Kelecioğlu, Trans.). Ankara: Nobel Publications

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16* (1), 5-18).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Erdogan, E. (2018). *The relationship between socioeconomic characteristics and reading skills of students in the international student assessment program.* (Unpublished Master's Thesis. Trakya University Institute of Social Sciences).

Erdogan, E., & Güvendir, M. A. (2019). The relationship between socioeconomic characteristics and reading skills of students in the international student assessment program. *Eskişehir Osmangazi University Journal of Social Sciences, 20,* 493-523.

Ferhan, M. (2018). *Psychometric properties of PISA 2012 mathematics interest scale according to classical test theory and item response theory* (Master's thesis, Hasan Kalyoncu University).

Freitas, P., Nunes, L. C., Reis, A. B., Seabra, C. & Ferro, A. (2015). Correcting for sample problems in PISA and the improvement in Portuguese students' performance. *Assessment in Education Principles Policy and Practice 23(4):*1-17. DOI: 10.1080/0969594X.2015.1105784

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Research in Personality, 29(2),* 168-188.

Gelbal, S. (1994). A comparison on P item difficulty index and b parameter of Rasch model and ability measures based on them. *Hacettepe University Faculty of Education Journal. (10),* 85-9.

Grek, S. (2012). "What PISA Knows and Can Do: Studying the Role of National Actors in the Making of PISA." *European Educational Research Journal. 11 (2*): 243–254.

Gul Ince, F. (2016). *Comparative analysis of TIMSS 2011 mathematics subtest item parameters according to CTT and MTK with bilog mg, multilog and r programs.* (Master's thesis, Gazi University, Ankara).

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment, 10(3),* 229-244.

Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Application.* Boston: Kluwer Academic Publishers Group.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publication.

Karasar, N. (2014). *Scientific research method* (26th Edition). Ankara: Nobel.

Kline, T. J. B. (2005). *Psychological testing. A practical approach to design and evaluation*. Sage.

Liou P.Y., & Hung Y.C. (2015). Statistical Techniques Utilized in Analyzing PISA and TIMSS Data in Science Education From 1996 To 2013: A methodological review. *International Journal of Science and Mathematics Education, 13(6),* 1449–1468.

Nartgün, Z. (2002). *Examining the item and scale properties of the Likert-type scale and the metric scale to measure the same attitude, according to the classical test theory and the theory of implicit characteristics.* (Unpublished doctoral dissertation, Hacettepe University, Ankara).

OECD (2004). *Learning for Tomorrow's World: First results from PISA 2003*. Paris: OECD Pub.

OECD (2007). *PISA 2006: Science competencies for tomorrow's World.* Paris: OECD Publishing.

OECD (2009). *PISA Data Analysis Manual*. Paris: OECD Publishing.

OECD (2010). *PISA 2009 Results: What students know and can do: Student performance in reading, mathematics and science.* Paris: OECD Publishing.

OECD (2013). *PISA 2012 Results: What students know and can do: Student performance in mathematics, reading and science.* Paris: OECD Publishing.

OECD (2020). PISA Participants. Web Address: https://www.oecd.org/pisa/aboutpisa/pisa-participants.html

Teacher, T. (1995). *Differential item functioning analysis of the verbal ability section of the first stage of the university entrance examination in Turkey* (Unpublished master's thesis). Middle East Technical University, Ankara.

Özdemir, C. (2016). Methodological review of studies using OECD PISA Turkey data. *Education Science Society, 14(56)*, 10-27.

Özer, M. (2020). What PISA tells us about performance of education systems? *Bartın University Journal of Faculty of Education, 9(2),* 217-228.

Pintrich, P., Smith, D. A. F., Garcia, T., & McKeachie, W. (1992). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Washington, DC: Office for Educational Research and Improvement.

Rajchert, J. M., Żułtak, T., & Smulczyk, M. (2014). Predicting literacy and its improvement in the Polish national extension of the PISA study: The role of intelligence, trait-and state anxiety, socio-economic status and school type. *Learning and Individual Differences, 33*, 1-11.

Rosseel, Y. (2012). Lavan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2).*

Rutkowski L., Gonzalez E., Joncas M., & Von D, M. (2010). International Large-Scale Assessment Data: Issues in secondary analysis and reporting, *Educational Researcher, 39*, (2), 142–151.

Shala, A., &. Grajcevci, A. (2018). Kosovo's low performance in PISA 2015: A study from a socioeconomic perspective. *Educational Process: International Journal, 7(1),* 48-59.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality in assessment and ability estimation. *Psychometrica, 55,* 293-325.

Strand, S. &. Demie, F. (2007). Pupil mobility, attainment and progress in secondary school, *Educational Studies, 33:3*, 313-331. DOI: 10.1080/03055690701423184

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. Northridge. Cal: Harper C.

Taş, U. E., Arıcı, Ö., Özarkan, H. B., & Liberty, B. (2016). International student assessment program *PISA 2015 National Report*. Ankara: Ministry of National Education http://odsgm.meb.gov.tr/test/analizler/docs/PISA/PISA2015

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural Assessment. *European Journal of Applied Physiology, 54(2004),* 119–135.

Waldow, F. (2013). PISA under examination: changing knowledge, changing tests, and changing schools. *Comparative Education, 49 (4),* 536-545.

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives, 30(3)*, 3-32.

Wood, P. (2008). Confirmatory Factor Ana. for Applied Res. *The American Statistician,62(1)*, 91–2.

Yalçın, S., & Tavşancıl, E. (2014). The comparison of Turkish students' PISA achievement levels via data development analysis. *Educational Sciences: Theory and Practice, 14(3),* 961- 968.

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30(3),*187–213.

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data.* (Doctoral Dissertation). University of Massachusetts, Amberst.