The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2017

Volume 6, Pages 14-23

ICEMST 2017: International Conference on Education in Mathematics, Science & Technology

GENERATION STUDY OF PISA MATHS PROFICIENCY LEVELS IN TURKISH 6TH GRADE STUDENTS

T. Oguz Basokcu
Ege University

Simge Ceylan
Ege University

Abstract: PISA is an international exam which aims to assess whether 15-year-old students are able to convert their academic outcomes into solving daily life issues as well as analyzing high level cognitive skills. PISA evaluates the outcomes through item-based skills classification constituted by IRT technique with the help of the samples gathered from each participant country. Skill classification is a grouping process which helps to interpret the proficiency of students at different points in accordance with the ranges described for each level. For the Maths proficiency level of classes gathered by this process increasing from 1 to 6 hierarchically: the ability to give the correct answer at Level 1 only when all related information is presented and questions are clearly explained is recognized, whereas it is more frequent to recognize the correct answer at Level 6 in which high level cognitive skills are used, necessary knowledge is organized and interpreted to solve the problem. Of all the OECD countries, 15.8% of China and 10% of Japan are at Level 1, which is 52% for Turkey. An experimental study is being pursued in an attempt to enhance the Maths literacy success of 6th grades by increasing the number of implementations in large-scale international exams with TUBITAK Research Project numbered 115K531. About 3200 students are included within the project as a longitudinal study. The equivalence of the tests to that of PISA has been assured. At this point, the study aims to determine whether the classifications made for PISA Turkey similarly range also in the younger age group, as well as aiming to find out whether the origin of the distinction between Turkey and other OECD countries in the higher levels begins at an earlier age. In Izmir province, 6th grade students who were determined randomly by the stratification method were subjected to tests that required multiple levels of thinking and represented 6th grade Maths subjects through test items in the form of multiple choice, true-false and open-ended. Plausible scores appropriate for PISA procedures and the cut points determined by using those scores and PISA standards were designated and proficiency levels were obtained. The proficiency levels of 6th grade students in the sample were specified with the help of this method. When the results of the study are analyzed in detail, it is clearly seen that the percentages described in the PISA 2015 Report show a similar distribution across the classrooms.

***Keywords*:** Maths proficiency, PISA, rasch model.

Introduction

The Programme for International Assessment (PISA), first implemented in 2000 by the Organization for Economic Co-operation and Development (OECD) which is an intergovernmental organization of industrialized countries, is an international assessment system that measures math literacy, science literacy and reading skills of 15-year-old students every 3 years. According to 2015 data, it covers 35 OECD countries, 37 partner countries and economies.

The first goal of education for all politicians around the world is to fully realize the potentials of their citizens and to enable them to develop their skills in accordance with changing world conditions. In this context, PISA results indicate much more than points or rankings. It is also the world's leading education benchmark used to assess the quality, equality and productivity of school systems. This provides governments and educators with

- Selection and peer-review under responsibility of the Organizing Committee of the conference

*Corresponding author: Oguz Basokcu- E-mail: oguzbasokcu@gmail.com

the definition of effective educational policies that can adapt their own values by specifying the characteristics of high-performance educational systems. (PISA, 2015)

In Turkey, it is seen that the success graph of PISA is below the expectations and decreasing gradually. Therefore, in this research, we examined the similarities and differences in the ability distribution of the Turkish sample in the PISA applications at earlier education levels. The measurements made and the findings obtained were compared with the countries in the upper row in PISA applications and OECD averages. In this way, it is aimed to generalize the level of success that Turkey has in PISA applications to other education levels and to reach the findings about the reasons for the achievement below the expected level.

About PISA

The OECD Programme for International Student Assessment (PISA) is an international large-scale study that focuses on the capabilities of 15-year-old students' math literacy, science literacy and reading skills. PISA defines mathematical literacy as follows: An individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen.

Starting from 2000, PISA is conducted every 3 years with a primary focus on one area for each cycle.

Table 1. PISA cycle topics by years

PISA Administration cycle

Assessment year	2000	2003	2006	2009	2012	2015
Subjects assessed	**Reading** Mathematics Science	Reading **Mathematics** Science Problem solving	Reading Mathematics **Science**	**Reading** Mathematics Science	Reading **Mathematics** Science Problem solving	Reading Mathematics **Science**

(OECD, 2009)

From 2000 to 2015, each cycle focused on a different area. As shown in Table 1, the focus was on reading in 2000, followed by mathematics and science, and in 2009, the cycle began again with Reading and repeated in the same order. Apart from this, problem solving was added in 2003 and 2012 as well.

Pisa Sampling

PISA has a two-stage formed layered and random sample design. The first layer is the schools and the second is the students in the schools. The first step in the sample is the need to identify the target population of PISA students. It is generally considered to be 15-year-olds, but more precisely, it represents a sample of the age group of 15 and between the completed months of + 3 and -3 and the age group of 16 and the completed months of -2 and +2. The size of the sample taken from each country was determined as at least 150 school samples and at least 4500 students.

Ability Estimation in PISA RASCH

PISA uses Rasch Model among IRT (*Item Response Theory*) models statistically while determining the levels of student abilities. IRT is an approach that provides mathematical models which can overcome the weaknesses of classical test theory. It is a growing theory that psychometricians tend to use it, especially because of the claim of "sample-independent substance parameter" estimation and "ability to test independently". On the other hand, because of the large number of models available, it is possible to apply FTC analyzes to different measurement results. It also makes it easier to make accurate inferences about individuals and test items by offering the ability to compare individual skill levels with difficulty levels of questions, since individuals can calibrate the ability parameters and the difficulty parameters of the items at the same scale level. It offers different models such as

logistic models with 1, 2, 3 and 4 parameters according to the number of MTK parameters; single and multi-dimensional models according to the number of dimensions; dual and multiple (multi-categorized) scoring models. The Rasch Model is defined as a single-parameter model because the item characteristics curves depend solely on item difficulty. In the three-parameter logistic model, the characteristics curves of the item depend on (*i*) the item difficulty parameter, (*ii*) item discrimination parameter and (*iii*) the "guess" parameter. This last parameter concerns the possibility of all students in the multiple choice test to answer the item correctly, no matter how difficult it is.

The Rasch Model is designed to create a symmetrical continuity with both item difficulty and student competence. Item difficulty and student competence are related to a logistical function. With this function, it is possible to calculate the likelihood that a student will correctly answer an item. Moreover, because of this possibility connection, it is not necessary to apply every item sequence to every student. If some anchor items are warranted, Rasch Model may create a scale with each item and every student. This last feature of the Rasch Model is one of the main reasons why it is fundamental in educational research and especially in PISA practice (Edition, 2009)

Rasch is able to describe student ability continuously using dichotomic data. With three basic principles, we can lay the groundwork for the construction of Rasch continuity. The first principle concerns item difficulties. Take, for example, two items consisting of two questions. We cannot compare difficulties for these two items if the patterns of responses given to items 1 and 2 are (0, 0) and (1, 1) (indicating 1 success and 0 failure). On the other hand, the response pattern obtained in (1, 0) and (0, 1) is informative in terms of comparison. If we assume that the response pattern in this way is 50 students (0, 1) and 10 students (1, 0), we can reach the result that the second item is easier than the first. In fact, 50 students responded incorrectly to the first item, the second correctly answered, and only 10 responded correctly to the first item and the second item incorrectly. This indicates that when one person correctly answers one of the two items, the probability that the correctly answered question in the second item is 5 times the probability that it is the first item. Therefore, it is easier to answer the second question correctly than the first one correctly. However, we should not ignore that the relative difficulty of the two items is independent of the student abilities.

The second principle concerns the identification of the reference point. In the Rasch Model, the unit of measurement is defined by the probability function, which includes the item difficulty and the parameters of the student's ability. For this reason, it has been accepted that only one reference point has to be defined. The most common reference point is the zero point of item difficulties. However, accepting a zero center in the student's ability can be used as another relative reference point.

The third principle emphasizes continuity. Continuity plays a role in the calculation of the relative difficulty of the items that are presented to different subpopulations in part. Suppose that the first item is given to all students and the second item is given only to low-ability students. Comparisons of the items will only be made on the lower populations studied, i.e. on the low-skilled student population. The difficulty of the two relative items will depend on the common subgroup of these students (Edition, 2009). Student scores can be calculated when item difficulties are placed in Rasch continuity.

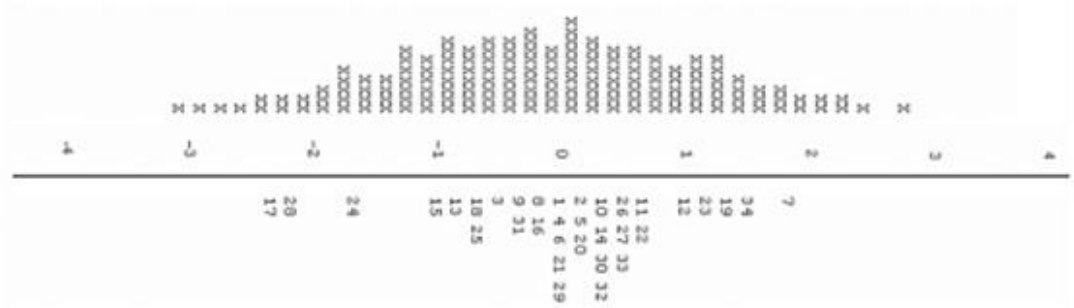

Figure 1. Student score and item difficulty distribution on a Rasch continuum

The line in Figure 1 represents Rasch continuity. Item difficulties are above and item numbers are below the line.

For example, item 7 represents a difficult item and item 17 represents an easy one. This test includes several easy items, a large number of intermediate items and a few difficult items. The symbols x above the line represent the distribution of student scores (OECD, 2009).

Calculating the student's score in Rasch Model

After the item difficulties are determined on the Rasch scale, student scores can be calculated. For a student whose ability is represented by B_i, the possibility of giving a correct answer to the item j whose difficulty level is represented by D_j is as follows:

$$p(X_{i,j} = 1 | \beta_i, \delta_i) = \frac{\exp(\beta_i - \delta_i)}{1 + \exp(\beta_i - \delta_i)}$$

Similarly, the possibility of giving a wrong answer is as follows:

$$p(X_{i,j} = 0 | \beta_i, \delta_i) = \frac{1}{1 + \exp(\beta_i - \delta_i)}$$

Rasch Model assumes the independence of items, so the probability of a correct answer is not dependent on the answers given to other items. As a result, the possibility of success from two items equals to the multiplication of two success probabilities.

Rasch ability estimations are often specified as the maximum likelihood estimation (or *MLE*). As shown in these figures, Rasch Model only returns a maximum probability estimate per raw score, i.e. zero correct answers, one correct answer, two correct answers, and so on.

Warm (1989) has shown that this maximum likelihood estimation is biased and suggested to weight the contribution of each item according to the information that this item can give. Warm estimations and *MLE*s are similar to students' individual skill estimations.

When the Warm estimation is corrected for the small bias in the MLE, it is usually an estimation of one's temperament. Therefore, in PISA, weighted likelihood estimations (*WLE*s) are calculated by applying weights to *MLE* in order to account for the bias inherent in *MLE*, as Warm proposed (OECD, 2009).

Plausible Value

Producing plausible values from a training test consists of drawing random numbers from posterior distributions. In its most basic sense, "Plausible values say that a learner is a demonstration of the abilities that it can have at a reasonable level. Instead of estimating the ability of a student directly, it estimates a student's probability distribution for Q. That is, instead of taking a point account for a Q as in *WLE*, a range of possible values of a student for Q and the combined probability for each of these values are estimated. Plausible values are random lines from this (estimated) distribution for a student's Q" (Wu and Adams, 2002).

All this methodology aims to create a continuum from a set of discontinuous variables (i.e. test score). It is aimed to avoid biased inferences as a consequence of measuring the underlying ability that cannot be observed through a test using a relatively small number of items.

Finally, an individual estimation of student ability can also be derived from posterior distributions. This derived individual estimation is called expected posteriori estimator (*EAP*). Instead of assigning a series of random values from the posterior distributions, the averages of the posterior distributions are given. For this reason, *EAP* can be considered as the average of a group of reasonable values for a particular student (Edition, 2009).

PISA Proficiency Levels

Proficiency levels have been proposed to be powerful tools that can be used to communicate results from large-scale assessment studies to the wider public with higher levels indicating higher proficiency. Importantly, proficiency levels describe the cognitive skills and the knowledge of which a student is capable (Fischbach, Keller, Preckel & Brunner, 2013).

Each proficiency scale is standardized to have M=500 and SD=100 across OECD countries. Furthermore, these scales can be subdivided into six proficiency levels for the mathematics and science tests and five proficiency levels for the reading test.

The first stage in creating proficiency levels begins with the putting possible scales and dimensions in written forms that can be used by the experts in each field for reporting. This step is defined as *identifying possible scales*. The advantage of this process is that multiple scales developed for the weighted area, which is concentrated in cycles, are more meaningful and potentially more useful for feedback and reporting purposes. The second stage deals with *assigning items to scales*. Each question item is associated with a thought scale. Experts, then, evaluate the properties of each item according to the classification in the evaluation framework. Then statistical analysis of the item scores obtained from the pilot application is used to obtain an objective criterion related to the distribution of the items in the scale. The skills are controlled in the third stage, which is known as *skills audit*. This stage involves analyzing the subject area of each item in detail by the expert, in relation to the definition of the relevant subscale in the evaluation framework, and evaluating the partial scores and points scored. The knowledge and skills required to achieve each point are described and explained. The fourth stage deals with *analyzing field trial data*. First, the data obtained from the pilot application is analyzed according to the IRT and the item difficulty for each achievement threshold is calculated. In general, when there is only one achievement bias for the items, more than one achievement threshold can be calculated for the ones that require partial scoring. Subsequently, achievement thresholds within each scale are placed along a continuous difficulty continuum, associated with student skills. The fifth stage includes *defining the dimensions*. The field expert combines the results from the analyzes done in stages 3 and 4. Subsequently, the item score steps for each set of scales are sorted by reference to the associated thresholds and then linked to the descriptions of the relevant knowledge and skills. These processes create a hierarchy of knowledge and skills that define the final dimension. The sixth stage consists of *revising and refining with main study data*. When this step is reached, the information obtained from the statistical analysis of the relative difficulty of the item thresholds is updated, as the data in the actual application is now ready to be used. After this, specialists are in charge of revising and checking the data. The seventh stage involves the *validating process*. First, knowledgeable experts are recruited who have the necessary materials to enable them to evaluate the indicators on which the levels defined for the PISA items are based. Then comes the consultation process during which the defined scales are presented to the national coordinators of all PISA countries. This stage allows one to reach the conclusions about how well the users of the defined levels find them informative (Anıl, Özkan & Demir, 2015).

PISA revises and updates description of proficiency levels each semester, which is determined in such a way as to reflect changes in evaluation and in the framework and requirements of new tasks developed for the evaluation. The most recent statement of proficiency levels is based on the PISA 2012 evaluation (OECD, 2014). PISA results demonstrate what is possible in education by showing what the students in the fastest growing education system can do best (PISA, 2015).

Content of PISA Mathematics Proficiencies

(Summary description of the six levels of mathematics proficiency in PISA 2015)

The mathematical proficiency levels identified by the PISA consist of six levels. While there is a hierarchical increase from 1 to 6 among these levels, the ability to respond correctly at the first level is given when all relevant information is provided and when the questions are clearly defined. Level 6, on the other hand, has the right frequency of answers where high-level thinking skills can be used, the information required for solution of the problem is organized, and where it is desired to be interpreted as the result. The item-based ability classification process is used when the score ranges of the levels are determined. Here, basically the difficulty levels of the items and the number of students who respond correctly to these items are taken as references. For this, the items are sorted according to their difficulty level. For instance, items 1 and 2 are in low difficulty, items 3 and 4 are in medium difficulty, and items 4 and 5 are in high difficulty. If the student cannot answer the 1st and 2nd items correctly, she is expected not to answer the 3rd and 4th items correctly as well. Assuming that another student can answer all the items correctly from 1 to 5, it is probably interpreted that she could answer the 6th item correctly. In the same way, it can be understood that a student who answered correctly the 1st and 2nd item but did not answer the 5th and 6th items correctly cannot answer the 4th item correctly, either. In Table 2, range of points are given by levels.

Table 2. Range of points for PISA 2012 mathematical proficiency levels

Under 1	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
- 357.77	357.77-420.07	420.07-482.38	482.38-544.68	544.68-606.99	606.99-669.30	+ 669.30

Aim

This research focuses on the distribution of skills of the Turkish sample in PISA applications, their similarities and differences in earlier education levels. The measurements made and the findings obtained were compared with the countries in the upper rows in PISA applications and OECD averages. In this way, it is aimed to generalize the level of success that Turkey has in PISA applications to other education levels and to reach the findings about the reasons for the achievement below the expected level.

Methods

Sample

The research population consists of 448 Secondary State Schools in 30 districts affiliated to İzmir Provincial National Education Directorate. There is a total of 1822 branches and 45069 students at the 6th grade level in these schools. The confidence level and the confidence interval statistics were used when determining the sample size (Oulte, 2011; Thompson, 2012). When the confidence level was set at 99% and the confidence interval was set at t=2, the sample size was set at n=3809 for the population of 45069 (Lodico, Spaulding & Voegtle, 2006).

Taking this sample size into consideration, the schools in İzmir province of Turkey were determined according to the districts by randomized cluster sampling method and 148 branches and 4592 students in 20 schools were included in the study for the research sample. With the final state of the sample size, the confidence interval of the sample has been reduced to t=1,77, meaning that the power to represent the population has been increased. The TUBITAK project, numbered 115K531, consists of a total of 2 experiments and 1 control group as it is an experimental and longitudinal study. This research includes only the experiment 1 and the experiment 2 groups of the project sample. Thus, the sample of this research consists of 2672 students in these two groups.

The Instrument

The instrument used in this study is the first of the 4 monitoring tests applied in the TUBITAK project, numbered 115K531. This test consists of two books and 11 items to measure high-level thinking skills. The statistics for the items are given in Table 3.

Table 2. Test item parameters

No	Question Code	Item Difficulty (Pj)	Item Discrimination (rbis)
1	9796	0.41	0.5
2	3690	0.21	0.22
3	1027	0.18	0.45
4	1025	0.27	0.55
5	1033	0.51	0.48
6	1021	0.68	0.45
7	1032	0.4	0.41
8	7339	0.01	0.19
9	6728	0.29	0.38
10	5158_B	0.38	0.52
11	5158_C	0.31	0.49
	Average	0.33	0.43

As shown in Table 3, the item difficulty values range from 0,10 to 0,78 and the item discrimination values from 0,15 to 0,81. The average item difficulty was 0,31, and the average item discrimination was calculated as 0,56. As a result of the analysis, it was found that the test had sufficient discriminative value. At the same time, with the help of pilot implementation, the scoring keys to be used for open ended questions for the test items were determined.

Table 4. Test descriptive statistics

Average	3.65
Median	3
Standard Deviation	2.07
Variance	4.28
Skewness	0.63
Kurtosis	0.02

As shown in Table 4. the average score of the test is 3.65 and median score is 3. The standard deviation of the test was 2.07 and the variance was 4.48. Skewness and kurtosis values were found to be 0.626 and 0.016. respectively. A value of less than 1 means that the distribution does not deviate too much from the normal distribution.

It is very important to prove that the questions developed in the project have the same level and psychometric properties as the PISA and TIMSS questions. To this end. the averages were taken so that the statistics of the questions general test can be calculated. ANOVA analysis was performed for repeated measures to determine whether the developed questions correspond to PISA and TIMSS questions. According to the analysis results. there is no statistically significant difference between PISA. TIMSS and the average of the project questions ($F_{(896-2)}$=2.358. p>0.5).

Whether the variances of distributions are equal or not is examined by the Mauchly Sphericity test. According to the results of the analysis. it was seen that the assumption of sphericity is not distorted. that is. the variances of distributions are equal. ($\chi^2_{(2)}$=4.881. p>0.5). This is an indication that the questions produced within the project correspond to the PISA and TIMSS questions. At the same time. however. correspondence has also been examined in terms of scope and criteria. Test scores according to the criterion were statistically significant and highly correlated with each other (r=0.52. p<0.01).

Scope validation work was carried out by a team of 5 people consisting of expert project researchers and consultants in the field of two measurement and evaluation. two mathematics education and one program development. with at least associate degree. As a result of the study. it is seen that the questions produced in the project are compatible with PISA and TIMMS coverage.

Results and Findings

The research aims to examine the distributions of the mathematical proficiency levels determined for the students in the PISA applications in Turkey and to compare the distributions determined in a similar measurement targeting an earlier period than the PISA age range. To this end. it is first necessary to examine how the levels of proficiency are distributed to countries and the general average. Some results and country-based comparisons based on PISA 2012 results are given in Table 5.

Table 5. Country comparisons for PISA 2012 mathematical proficiency levels. as percentages

	Shanghai	Finland	OECD Avr.	Turkey
Level 1⁻	0.85%	3.34%	8.02%	15.48%
Level 1	2.95%	8.92%	14.98%	26.50%
Level 2	7.51%	20.49%	22.46%	25.54%
Level 3	13.10%	28.82%	23.74%	16.52%
Level 4	20.17%	23.17%	18.15%	10.09%
Level 5	24.60%	11.71%	9.34%	4.67%
Level 6	30.83%	3.54%	3.31%	1.20%

Looking at the percentage distributions obtained with reference to the proficiency levels in Table 5. there is a 0.85% part in Shanghai below Level 1. while this corresponds to 8.02% in the OECD average. In Turkey. this ratio is 15.48%. When we look at Level 6. it is seen that only 1.20% of the students have reached this level in Turkey while there is a 30.83% part in Shanghai. The distributions of the countries can be seen more clearly in Figure 2.

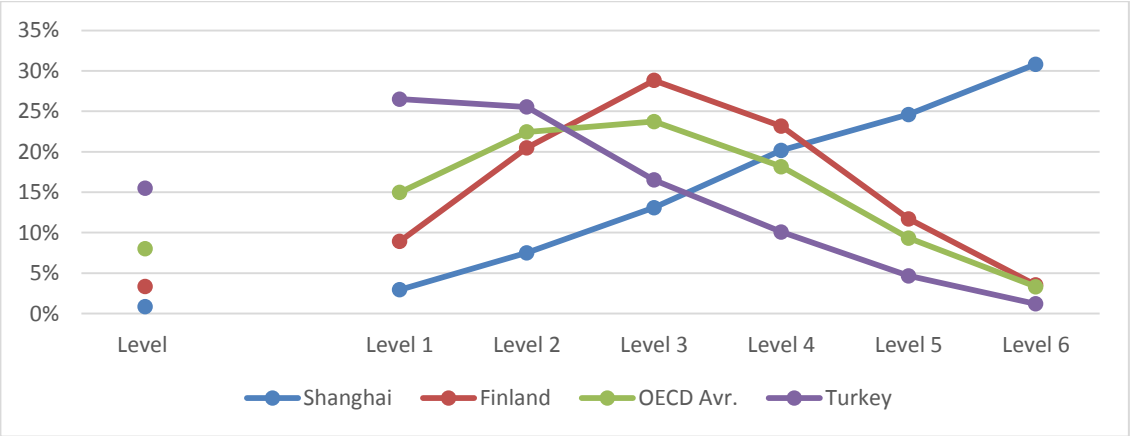

Figure 2. Country comparisons for PISA 2012 mathematical proficiency levels. as percentages

When the graph is examined in general. it is observed that Shanghai is distorted to the left. the average of Finland and OECD is normal. and Turkey is distorted to the right. This gives a good idea of the level of countries as a measure of success. Table 6 gives the mean and standard deviations of the compared countries.

Table 6. PISA 2012 Math average scores and standard errors by countries

	Shanghai	Finland	OECD Average	Turkey
Average Score	613	519	494	448
Standard Error	3.3	1.9	0.5	4.8

According to Table 6. Shanghai's average score is quite high when compared to the OECD average. When Turkey's average score is examined. it is seen that it is considerably lower than the other countries.

Within the scope of the project. the test scores proved to be corresponding to the PISA content and level were first analyzed with the Rasch Model in accordance with the PISA procedure. and the ability scores of the sample were determined and then statistically more reliable ability distribution was obtained by calculating plausible scores. Table 7 demonstrates the comparison of cut-off scores of percentage levels for PISA 2015 and Project measurements.

Table 7. PISA Turkey percentage cut-off scores for ages 15 and 12

	PISA Turkey Age 15	Project Age 12
10th	339	352
25th	382	392
50th	438	429
70th	507	499
90th	577	566
Mean	448	450

When Table 7 is examined. it is seen that the cut-off scores of 15-year-old and 12-year-old Turkey samples are very similar. Figure 3 shows the overlap level of cut-off scores for two tests.

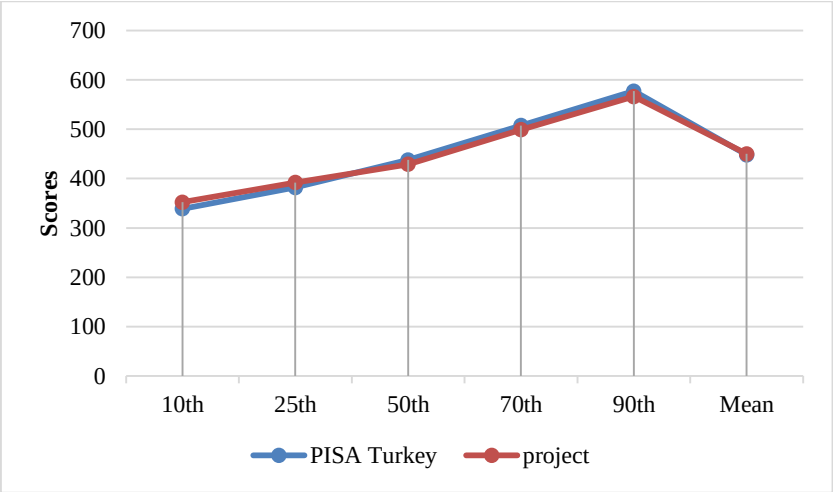

Figure 3. Overlap level of cut-off scores of PISA Turkey and the project

As can be seen in Figure 3. the cut-off scores of the two groups are very close. This shows that the percentage distributions of the two samples are very similar. At the next stage. the two groups were analyzed in terms of mathematical proficiency as percentages. Table 8 shows the mathematical proficiency levels of the 15-year and 12-year Turkey distribution.

Table 8. Percentage distribution of mathematical proficiency levels for PISA turkey and Project measurements

	PISA Turkey Age 15	Project Age 12
Level 1⁻	15.5	13.9
Level 1	26.5	18.6
Level 2	25.5	37.0
Level 3	16.5	20.2
Level 4	10.1	8.3
Level 5	4.7	1.4
Level 6	1.2	0.6

When Table 8 is examined. it is seen that the percentage distributions of mathematical proficiency levels are similar in both groups. In the age group of 15. Level 1⁻ is 15.5% while that of the 12-year-old is 13.9. Similarly. the 15-year-old group has a higher falling rate in Level 1. In Levels 2 and 3. the 12-year-old group appears to be in a higher percentage. At Level 4. 5 and 6. 15-year-old group is seen to be proportionally higher. Figure 4 shows the distribution of 12- and 15-year-old groups in terms of their mathematical proficiency levels.

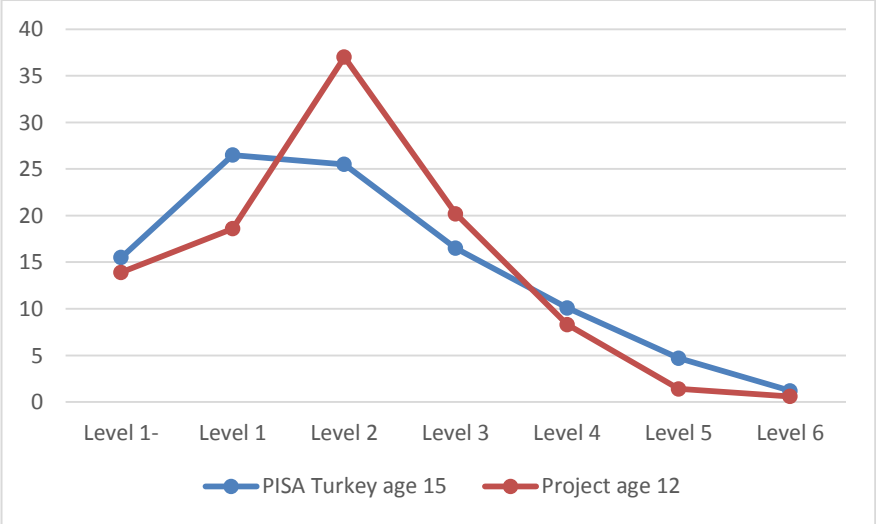

Figure 3. Distribution of 12- and 15-year-old groups in terms of their mathematical proficiency levels.

In Figure 4. it is seen that the 15-year-old group is higher at Level 1⁻ and Level 1. This indicates that in the sample of 15-year-olds. the number of students at Level 1⁻ and Level 1 is lower; that is. the number of students at the lower level is more than 12-year-old group. When Levels 2 and 3 are considered. it is seen that the

percentage distribution of 12-year-old group is higher in this range. It seems possible to say that the distribution for 12-year-old group is moderately more intense. At Levels 4. 5 and 6. the 15-year-old group is higher than the 12-year-old group. It was determined that there were some differences in the two groups. but these differences were not statistically significant. In this case. it is understood that students in the 12- and 15-year-old groups do not differ in terms of PISA mathematical proficiency levels.

Conclusion

As the research findings show. a distribution similar to the percentages described in the PISA 2015 report of proficiency levels was achieved by the project sample. There are great similarities between 6th grade students and 15-year-old group in terms of mathematical proficiency levels. More specifically. project sample and PISA Turkey measurements have very close values in terms of the mean and standard distribution as well as cut-off point scores. point values falling in percentages and sample percentages of proficiency levels.

Findings indicate that there is no difference between the levels of having high level mathematical proficiency levels in the 6th grade and the 15-year-old group in Turkey. Although the research includes test correspondence. sample validity. and psychometric properties of the questions. it is clear that additional research is needed to generalize the findings since only one province and a single class level are taken as basis.

It is also seen that during the international examinations (PISA. TIMSS. PEARLS). the low level of achievement of the students in Turkey shows similar features at earlier stages of the education system. In this sense. we can say that in order to increase the level of Turkey's success in international examinations. more holistic and systematic solutions are needed.

References

Anıl. D.. Özkan. Y. Ö. & Demir. R. E. (2015). *Pisa 2012 araştirmasi ulusal nihai rapor*.

Bodin. A. (2005). What does PISA really assess? What it doesn't? A French view 1. (June). 1–25.

Edition. S. (2009). *PISA Data Analysis Manual: SPSS. Second Edition*. *Analysis*. https://doi.org/10.1787/9789264056275-en

Fischbach. A.. Keller. U.. Preckel. F. & Brunner. M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*. *24*. 63–72. https://doi.org/10.1016/j.lindif.2012.10.012

Framework. R. (2013). 2 • Figure 2.2 •. 2012–2014.

Linneweber-Lammerskitten. H. & Wälti. B. (2005). Is the definition of mathematics as used in the pisa assessment framework applicable to the harmos project? *ZDM - International Journal on Mathematics Education*. *37*(5). 402–407. https://doi.org/10.1007/s11858-005-0028-y

OECD. (2009). *PISA Data Analysis Manual: SPSS*. *Oecd*. https://doi.org/10.1787/9789264056275-en

PISA. (2015). *PISA 2015 Results in Focus*. *Oecd* (Vol. I). https://doi.org/10.1787/9789264266490-en

Scale. P. (2014). Proficiency Scale. 1–4.

Tienken. C. H. (2017). Understanding PISA Results. *Kappa Delta Pi Record*. *53*(1). 6–8. https://doi.org/10.1080/00228958.2017.1264806

Wu. M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*. *31*(2–3). 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Wu. M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects*. *39*(1). 33–46. https://doi.org/10.1007/s11125-009-9109-y