ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 106-117

# A Comparison of Different Designs in Scoring of PISA 2009 Reading Open Ended Items According to Generalizability Theory

Meral ALKAN* Nuri DOĞAN**

**Abstract**

This study compares the different designs obtained through four raters' scoring the open-ended items used in PISA 2009 reading literacy altogether or alternately according to the Generalizability Theory. The sample of the research was composed of 362 students (out of 4996 students participating in PISA 2009) who responded to the items of reading skills and who were scored by more than one rater. Two designs were created so as to be used in generalizability theory in the study. One of them was the crossed design symbolized as "s x i x r" (student x item x rater), in which students are scored by each rater in terms of the same skills. The second was the nested design symbolized as "(r:s) x i", where each rater scored only a group of students and raters are nested in students and the items were crossed with these variables. On comparing the s x i x r design with (r:s) x i design, it was found that the relative and absolute error variances estimated for (r:s) x i design were smaller than those for s x i x r design and that therefore the G and Phi coefficients took on bigger values. On increasing the number of raters in both designs, the G and Phi coefficients also increased in the D study. While acceptable values of G and Phi coefficients were reached on reducing the number of raters by half in Booklet 2, raising the number of raters seemed more appropriate in Booklet 8.

*Keywords: Generalizability theory, reliability, G study, D study, PISA 2009*

## Introduction

At the beginning of the 21st century, social, economic, and technological developments have caused rapid change in every field. The desire of societies to keep up with this change has brought the issue of the quality of education to the fore. The quality of education is the most important factor in equipping new generations with new skills and competencies to keep up with this change. In today's world, where knowledge is accepted as a power and spreads rapidly, raising individuals who think critically, question, are responsible for their own learning, creative, and ready for life has become the most important goal of education systems. This situation affected assessment practices as well as educational practices. If subject knowledge alone is not a sufficient criterion, tests based on choosing the correct answer among the given options are not sufficient on their own. This understanding brought to learning has revealed the necessity of organizing the tests in a form in which the individual can structure their own answers and the curriculum from low-level thinking to an understanding that requires high-level thinking; teaching methods and techniques from a teacher-centered structure to a student-centered structure; assessment and evaluation approaches, on the other hand, have transformed from a structure that measures the extent to which information is acquired, to a structure that measures how information can be used in new situations or in real life (Biemer, 1993). This situation has been the trigger for turning to different approaches in the teaching process. OECD PISA (Programme for International Student Assessment) tests are designed to assess how well students, at the end of compulsory education, can

* Lect. PhD., Gazi University, Rectorate, Ankara-Türkiye, meralalkan@gazi.edu.tr, ORCID ID: 0000-0001-9497-3660
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Alkan, M., & Doğan N. (2023). A comparison of different designs in scoring of PISA 2009 reading open ended items according to generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 106-117. https://doi.org/10.21031/epod.1210917

*Received: 28.11.2022*
*Accepted: 12.06.2023*

apply their knowledge to real-life situations and can, therefore, fully participate in society (OECD, 2017; EARGED, 2010), and Turkey has been taking part in PISA and other international studies such as TIMSS and PIRLS assessing students' achievement comparatively on an international scale since the early 2000s. Turkish students cannot attain the desired success level in open-ended items in those studies (Balbağ, Leblebicier, Karaer, Sarıkahya & Erkan, 2016). The type of questions in which Turkish students attained the highest percentage of success level was multiple choice items in the sub-fields of reading skills and science literacy in PISA (Demir, 2010). Yet, open-ended items are considered more appropriate for measuring students' upper-level thinking skills (Ministry of National Education, 2017). For this reason, in parallel to the developments in the world, there has recently been a tendency to use open-ended items in selection and placement tests administered in Turkey in the transition both into secondary education and higher education. The Measuring, Selection and Placement Center (ÖSYM) made an announcement about the intention of using open-ended items in the Undergraduate Placement Exam (LYS) and published sample open-ended items regarding those would be used in the exam (ÖSYM, 2017). However, it is observed that multiple-choice tests are used more frequently than other types of tests in most institutions and establishments of education, in teachers' in-class activities, in student selection examinations, and especially in assessments involving a great number of students. The major reason for this is that using multiple-choice tests has certain advantages. Probably, one of the most important advantages is that scoring the correct answers does not change from rater to rater since answering the items requires only choosing one of the given options and there is no need to determine the degree of accuracy in scoring. In these tests, scoring consists of counting the correct answers and is purely objective (Özçelik, 2010). The greatest restriction of multiple choice tests is that they cannot measure high levels of performance, abilities, or skills (Konak, 2010; Güler, 2013). Classical testing methods such as multiple-choice, short answer, True-False, matching, and fill-in-the-blank used in assessing students' behaviours are incapable of determining upper-level mental processes such as problem-solving, reading comprehension, critical thinking, analytical thinking, empathising, researching, decision-making, understanding the importance of social history, and creativity (Kutlu, 2006). Open-ended items, on the other hand, enable students to create their own answers, give different personal answers, and answer the items from their own perspective.

When more than one rater is used in the assessment process, they can be the source of variance; they can cause errors and thus reduce the reliability of the assessment. Error components in individuals' scores are due to a variety of factors that introduce measurement error into the scores, such as intraindividual factors, characteristics of the measure itself, administration factors, scoring errors, and so on (Goodwin, 2001). The crucial concern about performance assessment and scoring of open-ended items is the objectivity of scoring as it is not easy to assess performance objectively, unlike traditional assessments (e.g., fixed response items) (Romagnano, 2001). And in the scoring of the open-ended items, the different factors interfere with scoring and reduce reliability (Atılgan et all, 2011). Rater reliability is the consistency between the scores given to a certain property. There are biases in scoring arising from raters. Scullen et al., (2000) describe raters' influence as "a broad category of effects which are not related to students' real performance but are related directly to raters who cause systematic errors in performance evaluation". It may be said that raters' influence arises from such psychological states as motivation, anxiety, achievement, and self-efficacy (Bernardin & Villanova, 2005), from personal traits (Wexley & Youtz, 1985), from their former beliefs, demographic properties such as gender and age, and raters' experience in scoring (Weigle, 1998). Raters' interaction with other sources of variability mingled in measurement is also important in reliability (Brennan, 2001). Therefore, errors arising from several sources of variability should also be taken into consideration in determining reliability. Error and error sources involved in measurement results should be well defined and methods should be found to estimate the amount of error (Turgut, 1992). The accuracy of the measurement results is very important as it affects the decisions to be made based on these results. Reliability can be defined as the degree to which measurement results are free from random errors (Baykul, 2000). One of the methods capable of analysing different sources of variability and the interactions between those sources altogether is Generalizability Theory (G Theory) (Shavelson & Webb, 1991). Due to the fact that G Theory considers more than one source of errors at the same time, it is thought that analysing the international studies in which Turkey also takes part from this perspective would be beneficial. In the

study conducted by Goodwin (2001), 3 approaches used in the inter-rater concordance and reliability study were compared: a) Percent of Agreement and Kappa b) Simple Correlation Methods c) G Theory techniques. In the study, 10 students were scored by 2 raters for the quality of their physical activities on 6 different days. Each rater evaluated the students over 7 points (1-lowest, 7-highest). As a result of the comparison of the advantages and disadvantages of different approaches, it is emphasized that the G Theory techniques are the most comprehensive, most flexible ones and allow the isolation of measurement errors caused by different sources in a study. It has been stated that the G Theory is the approach that gives the most information about the generalizability or reliability of the scores. Lee (2005), in his study, investigated the effect of the change in the number of tasks and raters in the TOEFL test on generalizability and tried to determine the most appropriate number of tasks for maximum reliability. As a result of the study, it was seen that increasing the number of tasks was more effective than increasing the number of raters. Therefore, it was concluded that using fewer raters in performance evaluation is appropriate for an acceptable level of generalizability. In a study on the generalizability of TIMMS open-ended items, Smith (1997) analyses the effects of raters. The researcher analyses the answers given by 150 students to each item with 50 booklets in each of seven English-speaking countries. It is stated that the number of raters should be raised to 15 from 5 for a generalizability level of 0.80 in all items and that this situation can cause problems in countries where the number of raters is small. Brennan (2001) states that the proficiency criteria of the reliability coefficients vary voluntarily, but some researchers may consider it "high" if the G and Phi coefficients are greater than 0.80. Sharma and Weathers (2003) investigated the generalizability of scales used in international research projects in all participating countries. It was concluded that the scale had the same meaning in all countries and that it was not specific to a certain country. It was also concluded that if the level of generalizability was 0.90, using 11 items out of 17 was sufficient for that level. Using a minimum number of items decreases examination/questionnaire time, eliminates the effects of tiredness and lowers costs. Thus, it is thought that knowing the adequate number of raters will be used in PISA would help to reduce labour and costs.

The number of raters to fulfill the task of scoring open-ended items as well as scoring time increases depending on the number of students participating and the number of items to be scored. For instance, 16 Turkish / Turkish Language and Literature teachers were needed only in scoring the reading literacy items of PISA 2009 (OECD, 2012). Appointing teachers to the task of scoring causes problems since schools are still open during this period and teachers have teaching tasks in their schools as well.

Countries can form a combination of differing designs in PISA and a design can be applied if it is accepted by the PISA consortium (OECD, 2012). Therefore, it is expected that determining which of the designs would be more appropriate for use and determining the minimum number of raters to attain the desired level of generalizability in each booklet will contribute to such examinations in terms of time, costs, and labour.

This study is believed to set a model in determining inter-rater reliability for open-ended items in such international studies as TIMSS, PIRLS, and ICILS and in terms of performing decision studies and to shed light on studies to be conducted in the future.

Assessment, Selection and Placement Centre (ÖSYM) started to make use of open-ended items in the selection and placement examinations (ÖSYM, 2013, 2017). Under the title of *"Information about Open-Ended Items and Examples"* details were given. Although it was expressed as an open-ended item, it was seen that the question type mentioned was short-answer questions and it was stated that the answer will consist of a word, a number, or a sentence (ÖSYM, 2017). However, there are no reliability and decision studies for such items; besides, such issues as how many items would be adequate in those examinations in which there is a great number of participants and how many raters should score and in what design they should perform the task are of great importance. For this reason, it is believed that this study will function as a guide. This study compares the results of the G study obtained from the designs created through more than one rater's scoring students' reading skills in PISA 2009 altogether or alternately according to the G Theory with the results of the decision study conducted with those designs.

## Method

This is a case study since it determines the properties of scoring PISA 2009 reading skills, and it is descriptive. Descriptive studies are the studies describing an existing event or the properties of an individual or a group as it is/as they are and describing a current state quantitatively or qualitatively (Karasar, 1998).

## Population and Sample

The sample of Turkey in PISA 2009 was composed of 4996 students from 170 schools who were randomly chosen by PISA international center by stratifying them according to 12 statistical region classification (IBBS, NUTS) and school types. 362 students who answered the reading skills items in PISA 2009 and whose booklets were exposed to multiple scoring constituted the sub-sample.

For the main survey, it was recommended to have 16 coders to code reading, 8 coders to code mathematics, and an additional 8 coders to code science items. Other possible coding designs were 16 reading and 8 mathematics and science coders or 16 reading, 4 mathematics, and 4 science coders. These numbers of coders were considered to be adequate for countries testing between 4 500 (the minimum number required) and 6 000 students to meet the timeline of submitting their data within 3 months of testing (OECD, 2012).

National Project Managers (NPMs) were responsible for recruiting appropriately qualified people to carry out the single and multiple coding of the test booklets. It was not necessary for coders to have high-level academic qualifications, but they needed to have a good understanding of the language of the test and to be familiar with ways in which secondary-level students express themselves.

In Turkey, the population of raters was composed of Turkish or literature teachers who had experience in international projects as a rater before or who had been teaching 15-year-old students (from $7^{th}$ graders to $10^{th}$). An important factor in recruiting coders was that they could commit their time to the project for the duration of the coding, which was expected to take up to one month. An official letter has been sent to schools to identify potential teachers with these qualifications. As a result of this process, 16 teachers were selected to take part in the scoring.

From 13 booklets used in PISA 2009, booklets 2 and 8 which contained reading skills and scored by four raters were used in this study. Booklet 2 contained six items whereas booklet 8 contained eight items.

## Research Data

The data collected from multiple raters scoring of reading skills in PISA 2009 constituted the data of this study. The data were provided by the Educational Research and Development Directorate (EARGED).

Two designs were created so as to be used in G Theory in the study. One of them was the crossed design symbolized as "s x i x r" (student x item x rater), in which students were scored by each rater in terms of the same skills. The second was the nested design symbolized as "(r:s) x i", where each rater scored only a group of students and raters nested in students, and the items were crossed with these variables.

## Data Analysis

This study aims to compare the results of the generalizability (G) and decision (D) studies of the scores of reading literacy items in PISA 2009 according to the crossed (s x i x r) and nested ((r:s) x i) designs and to compare the G and Phi coefficients as estimated by increasing or decreasing the number of raters in these designs. The data were analysed on the basis of these designs.

EDUG 6 programme was used in estimating variance components of the designs through G Theory, in calculating the rates of explaining the total variance of variables, and in performing the decision study for each design. EDUG 6 programme was developed for G Theory analyses, and it enables researchers to perform G and D studies for sources of variability they describe and for the designs they form with those sources of variability.

## Findings and Interpretations

In this section, the variance components and percentages explaining the total variance for crossed (s x i x r) and nested ((r:s) x i) designs and Generalizability levels and the results for D Study performed by changing the number of raters in these designs in Booklet 2 and Booklet 8 will be given.

**Table 1**

*Variance Components for "s x i x r" and "(r:s) x i" Designs and Percentages Explaining the Total Variance in Booklet 2 and Booklet 8*

| | Crossed Design | | | | | | Nested Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sources of variance | Squares total | Degrees of freedom | Squares average | Variance | % | Sources of variance | Squares total | Degrees of freedom | Squares average | Variance | % |
| **Booklet 2** | Students | 267.59 | 99 | 2.70 | 0.10577 | 16.0 | Students | 287.68 | 99 | 2.90 | 0.11076 | 18.5 |
| | Items | 137.14 | 5 | 27.42 | -0.00095 | 0.0 | Items | 78.79 | 5 | 15.75 | 0.03899 | 6.5 |
| | Raters | 39.53 | 3 | 13.17 | -0.02470 | 0.0 | r:s | 156.45 | 300 | 0.52 | 0.01439 | 2.4 |
| | si | 60.85 | 495 | 0.122 | -0.03699 | 0.0 | si | 79.91 | 495 | 0.16 | -0.06844 | 0.0 |
| | sr | 92.80 | 297 | 0.312 | 0.00692 | 1.0 | | | | | | |
| | ir | 419.35 | 15 | 27.95 | 0.27686 | 41.9 | | | | | | |
| | sir, e | 402.31 | 1485 | 0.27 | 0.27092 | 41.0 | ir:s, e | 652.79 | 1500 | 0.43 | 0.43519 | 72.6 |
| | Total | 1419.59 | 2399 | | | 100% | Total | 1255.64 | 2399 | | | 100% |
| **Booklet 8** | Student | 398.14 | 99 | 4.02 | 0.09072 | 19.8 | Students | 434.02 | 99 | 4.38 | 0.07488 | 13.7 |
| | Items | 14.25 | 7 | 2.03 | 0.00039 | 0.1 | Items | 14.99 | 7 | 2.14 | 0.00484 | 0.9 |
| | Raters | 60.97 | 3 | 20.32 | 0.02181 | 4.8 | r:s | 607.90 | 300 | 2.02 | 0.22295 | 40.9 |
| | si | 114.08 | 693 | 0.16 | -0.01120 | 0.0 | si | 141.53 | 693 | 0.20 | -0.00964 | 0.0 |
| | sr | 345.49 | 297 | 1.16 | 0.11923 | 26.0 | | | | | | |
| | ir | 40.41 | 21 | 1.92 | 0.01715 | 3.7 | | | | | | |
| | sir, e | 435.36 | 2079 | 0.20 | 0.20941 | 45.7 | ir:s, e | 509.84 | 2100 | 0.24 | 0.24278 | 44.5 |
| | Total | 1408.73 | 3199 | | | 100% | Total | 1708.30 | 3199 | | | 100% |

Variances estimated through G study and percentages explaining the total variance in Booklet 2 and Booklet 8 are given in Table 1. The variance component of the variable of students in Booklet 2 explains 16% of the total variance for crossed design while it explains 18.5% for nested design. We can see almost the same pattern in Booklet 8 and the variance component of the variable of students explains 19.8% of the total variance in crossed design whereas it explains 13.7% in nested design in this booklet. The variance component of students indicates that students differ in terms of reading skills. This pattern is similar in both designs and in both Booklets. In generalizability studies, variance due to students is considered as a universe score and this variance shows the difference between students in terms of characteristic which was measured (Brennan, 2001; Shavelson & Webb, 1991).

Accordingly, the percentage of the variance components of items explain the total variance is 0% for crossed design while it is 6.5% for nested design in Booklet 2 and the percentage of the variance components of items explain 0.1% of total variance in crossed design while it explains 0.9% in nested design in Booklet 8. It is clear in this case that items do not differ in terms of difficulty in crossed design but that they differ in nested design in Booklet 2. The fact that variance components are bigger in nested design is indicative of the fact that tasks are discriminated better and items do not differ in terms of difficulty in both designs in Booklet 8.

In Booklet 2, it may be stated that the variance component of raters' influence is quite small in crossed design and that therefore students are scored consistently in this booklet. In Booklet 8, it may be said that the value is high and that students' scores differed from one rater to another in crossed design. Since raters are nested within students, it is impossible to separate the raters' main effect from the interaction between students and raters. We interpret the substantial variance component for those combined effects (r:s=0.01439; 2.4 % of the total variance) in Booklet 2, and (r:s=0.22295; 40.9% of the total variance) in Booklet 8 as indicating student behavior differed from one rater to another. We do not know whether one rater produced more behavior than another (rater main effect), whether the relative standing of the student differed from one rater to another (student-by-rater interaction), or both (Shavelson & Webb, 1991. On examining the joint item rater variance component in Booklet 2, it is clear that the variance component is small in value and that raters do not differ in scoring from one item to another.

The residual variance was found to be high in both designs in Booklet 2 (sir,e=0.27092; 41.0% of the total variance; ir:s,e=0.43519; 72.6% of the total variance) but it was higher in the nested design. And also in Booklet 8, the residual variance was found to be high in both designs (sir,e=0.20941; 45.7% of the total variance; ir:s,e=0.24278; 44.5% of the total variance). For the variance component obtained from the interaction of three sources of variability to be zero (0) is a desired situation. The large residual component indicates that a substantial amount of variation is due to these confounded sources of variation (Shavelson & Webb, 1991).

**Table 2**

*Generalizability Levels for s x i x r and (r:s) x i Designs*

| | Crossed Design | | | Nested Design | |
|---|---|---|---|---|---|
| Booklet 2 | G coefficient | 0.89 | Booklet 2 | G coefficient | 0.99 |
| | Phi coefficient | 0.81 | | Phi coefficient | 0.98 |
| Booklet 8 | G coefficient | 0.71 | Booklet 8 | G coefficient | 0.95 |
| | Phi coefficient | 0.68 | | Phi coefficient | 0.85 |

On comparing the "s x i x r" and "(r:s) x i" designs, it was found that the G and Phi coefficients obtained from the (r:s) x i design were higher than those obtained for the s x i x r design. Having a generalizability coefficient of >.80 is a desirable situation (Mushquash & O' Connor, 2006). It was found in crossed design results, especially for booklet 8, that the G coefficient was not at an acceptable level and that the Phi coefficient was also below that level. However, the G and Phi coefficients for the same booklet were 0.95 and 0.85 respectively in the nested design and thus they were above the acceptable level.

**Table 3**

*Results for D Study Performed by Changing the Number of Raters in the "s x i x r" and "(r:s)xi" Designs*

| | Design | Crossed Design | | | | | Nested Design | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rater | 2 | 3 | 4 | 5 | 6 | 2 | 3 | 4 | 5 | 6 |
| Booklet 2 | G | 0.80 | 0.86 | 0.89 | 0.91 | 0.92 | 0.95 | 0.96 | 0.97 | 0.98 | 0.98 |
| | Phi | 0.68 | 0.76 | 0.81 | 0.84 | 0.87 | 0.92 | 0.94 | 0.95 | 0.95 | 0.95 |
| Booklet 8 | G | 0.56 | 0.65 | 0.71 | 0.76 | 0.79 | 0.80 | 0.86 | 0.89 | 0.91 | 0.92 |
| | Phi | 0.52 | 0.62 | 0.68 | 0.73 | 0.76 | 0.61 | 0.68 | 0.72 | 0.74 | 0.76 |

According to Table 4, the G coefficient obtained through scoring 100 students by 4 raters in terms of 6 items in booklet 2 in the crossed design is 0.89 and the Phi coefficient is 0.81. In nested design, on the other hand, the G coefficient in scoring 100 students by 4 raters is 0.97 and the Phi coefficient is 0.95- which are high values. On reducing the number of raters to 2 for scoring 100 students the G coefficient is found to be 0.80 and the Phi coefficient is found to be 0.68 in crossed design. However, the G and Phi coefficient were 0.95 and 0.92 respectively in the nested design.

Accordingly, the G coefficient found by scoring 100 students by 4 raters in terms of the 8 items in booklet 8 is 0.71 and the Phi coefficient is 0.68 in crossed design. This level of generalizability is well below 0.80- which is the acceptable level. The G coefficient with the same number of raters scoring the same number of students is 0.89 and the Phi coefficient is 0.72 in nested design.

On reducing the number of raters to 2 in scoring 100 students in crossed design, the G and Phi coefficients were found as 0.56 and 0.52, respectively. Yet, the G coefficient was 0.80 and the Phi coefficient was 0.61 in the nested design.

## Discussion and Conclusions

### Discussion

In the literature, G Thoery techniques are considered the most comprehensive and flexible in the estimation of interrater agreement and reliability (Goodwin, 2001) and G Theory is one of the methods capable of analysing different sources of variability and interactions between those sources altogether. In this study, the aim was to compare different designs (crossed and nested) according to G Theory and to find out the most effective way of scoring PISA open-ended items. Almost 5000 students participated in PISA 2009 assessment and the scoring process took almost a month to complete with the participation of 16 teachers as the raters. Reducing the cost and labour was the starting point of this study.

## Conclusion

On examining the results for Generalizability study in the s x i x r design for booklets 2 and 8 in PISA 2009, it was found that students differed in terms of reading skills in booklet 2, that items did not differ in terms of difficulty, that students' performance did not differ from item to item, that raters made a consistent assessment, that the student-rater interaction was very low and that raters' assessment did not differ from student to student. Yet, the joint effect of items and rater was very high. In this case, raters' scoring could be said to change from item to item.

It was found that raters differed in booklet 8 and that the joint effect of students and raters explained 26% of the total variance. Thus, raters' scoring changed from student to student. This situation manifested itself in the generalizability coefficient and the level of generalizability remained at 0.71.

However, the variance components for the joint effect of student item and rater explained the total variance at a high percentage. For the variance component obtained from the interaction of three sources of variability to be zero (0) is a desired situation. Having this ratio high can indicate that the student, item, rater interaction, and/or sources of random error can be big. On comparing the s x i x r and (r:s) x i designs, it was found that the relative and the absolute error variances estimated in the (r:s) x i design was smaller than in the s x i x r design, and thus the G and the Phi coefficients took on bigger values.

On examining the decision studies performed in both designs, it was found that increasing the number of raters provided an increase in the G and Phi coefficients in both designs but that the increase in the G coefficient obtained in this way was not big enough to bring advantages in terms of being economical. It was found that it was possible to reach acceptable levels of G coefficient by reducing the number of raters by half in Booklet 2.

The G and Phi coefficients were calculated as 0.89 and 0.81 respectively in booklet 2. When the number of raters is 5 and the number of students is kept constant, the G coefficient was calculated as 0.91 and the Phi coefficient as 0.84. When the number of raters is 6, the G coefficient was 0.92 and the Phi coefficient was 0.86. On increasing the number of raters, there was an increase in the G and Phi coefficients. But it was not substantial. When the number of raters was reduced to 3, the G and Phi coefficients were found to be 0.85 and 0.76 respectively whereas the G coefficient fell down to 0.80 and the Phi coefficient to 0.68 on reducing the number of raters to 2. In this situation, the Phi coefficient fell below the acceptable level while the generalizability coefficient remained at an acceptable level.

In booklet 8, the G and Phi coefficients were calculated as 0.71 and 0.68, respectively. This was below 0.80- which is the acceptable level for the G and the Phi coefficients (Mushquash & O' Connor, 2006). The Generalizability coefficient falls down to 0.56 and the Phi coefficient to 0.52 on reducing the number of raters to 2 from 4 and keeping the number of students constant. When the number of raters is 5, the G coefficient is 0.76 and the Phi coefficient is 0.73. When the number of raters is 6, the G coefficient is 0.79 and the Phi coefficient is 0.76. An increase occurred in the G and the Phi coefficients when we increased the number of raters. Yet, at least 6 raters are required to get a G coefficient at an acceptable level. It was found that the results obtained in earlier studies concerning G Theory were supportive of the ones obtained in this study. Different designs of G Theory were compared and the most appropriate number of items and the most appropriate number of raters were considered. Increasing the number of raters led to an increase in the G coefficient, but the effects of the increase diminished after a certain number of raters. Atılgan (2008), in a study concerning the generalizability of tests for selecting students in Music Department at Inönü University, found that it would be more appropriate to continue with the initial number of raters due to the fact that the increase in the G coefficient was not very effective. Smith (1997), in a generalizability study concerning the effect of the number of raters in the scoring of the open-ended items in TIMSS, found that the effect of the increase in the number of raters differs from one item to another and it would be more appropriate to raise the number of raters to 15 from 5 for the desired level of generalizability in all items and this shows that there may be a problem in countries where the number of raters is low. In some studies, it was found that increasing the number of tasks rather than raters was more efficient to maximize the score reliability. In a study concerning

generalizability of a performance assessment measuring achievement in eighth-grade Mathematics, Mcbee and Barnes (1998) investigated the effect of task similarity to generalize the results of the assessments. The Generalizability study results showed that the number of tasks required to reach acceptable levels of generalizability would be prohibitively high, even using only highly similar tasks. Schoonen (2005) found out that the generalizability of writing scores and the effects of raters and topics are very much dependent on the way the essays are scored and the trait that is scored. The overall picture is that writing tasks contribute more to the score variance than raters do. Lee (2005), in a generalizability study concerning the effect of number of raters and the number of tasks in the scoring of TOEFL writing assessment, reported that increasing the number of tasks rather than the number of raters per task would be more efficient to maximize the score reliability for writing. In a study of generalizability of students writing across multiple tasks, Hathcoat and Penn (2012) found that 77% of error variance may be attributable to differences within people across multiple writing assignments. D studies indicated that substantive improvements in reliability may be gained by increasing the number of assignments, as opposed to increasing the number of raters. Therefore, it was concluded that using fewer raters in performance evaluation is appropriate for an acceptable level of generalizability. In a study called the comparison of different designs in accordance with the generalizability theory in communication skills, Nalbantoğlu and Gelbal (2011) did G and D studies and compared crossed (s x t x r) and nested ((s:r) x t ) designs and observed that the variance that were estimated for variables in both designs were similar to each other and also D studies yielded the similar results and it was found that the scoring of certain number of students alternately (nested design) is much more convenient in time, labor, and cost. Polat and Turhan (2021) compared crossed and nested designs in language testing. G and Phi values and the variance associated with the student's main effect were higher, while the variance value of the residual effect was lower in crossed design. This study revealed that crossed designs could generate more reliable results in speaking exams. Zorba (2020) compared the result of a written exam used in personnel recruitment with different patterns in the generalizability theory. In this study, G and Phi coefficients were calculated as 0,33 and 0,29 for (p x i x r) (person x item x rater) design, and 0,76 and 0,64 for (p x (i : r )) (person x (item : rater) design, respectively. According to the results of the D study, it was observed that increasing the number of raters in crossed (p x i x r) design and increasing the number of items in nested (p x (i: r)) design increased the reliability. Khodi (2021), in a study of G-theory analysis of rater, task, and scoring method contribution, found that at least four raters (with G-coefficient = 0.80) were necessary for a valid and reliable assessment and he suggested student performance should be rated on at least two scoring methods by at least four raters.

Therefore, it is believed that determining the minimum number of raters by considering the degree to which an increase in the G coefficient is influential in results will help to reduce labour and costs in making decisions about determining the number of raters.

## Recommendations

Using the designs in which a group of raters scores a group of students alternately instead of having all raters score all students will be more economical in terms of time and labour if there is consistency between raters in performance determining examinations where there are a great number of students and more than one rater score them.

It was observed that a certain amount of decrease occurred in inter-rater consistency and in generalizability coefficients in partial scoring in the form of partial credit as 2-1-0 in booklets. Therefore, it should be made sure that a greater number of examples is given in training raters for booklets which are scored partially, the number of local examples should be increased, and scoring should be done on the item level not on the unit level. This means that all the items in a unit should be coded one by one and a new unit is started only after completing all the items in the previous unit in all booklets.

It is important that the group to function as raters in international activities was described beforehand. Assigning teachers for scoring from schools during school time and especially at the end of a semester for such activities lowers the teachers' motivation. Teachers should be able to perform the task of scoring with no fear of wasting time and disrupting their school work at the end of a semester.

Using open-ended items in a test for the transition from elementary education into secondary education and in a university entrance exam has been on the agenda. Using open-ended items in those examinations-which are extremely important in shaping students' future- is an issue that should be carefully worked on. It should not be forgotten that rater reliability is very important in assessing open-ended items. Considering the situations where the scoring of a booklet by 4 raters in PISA is inadequate, the number of staff to make assessment meeting the standards and the length of time required for the assessment of 2 million students and the reflections into the system should be calculated carefully.

Studies comparing different sources of variability (such as booklets, modules, etc.) and different designs in international performance assessment examinations such as TIMSS, PIRLS, PISA, and ICILS, in which Turkey takes part, could be performed.

Studies considering all the sub-fields such as mathematics literacy and science literacy in international performance assessment examinations such as TIMSS and PISA and analysing the correlations between them from the perspective of different scoring designs could be performed.

## Declarations

**Author Contribution:** Meral Alkan: Conceptualization, methodology, analysis, writing & editing, visualization. Nuri Doğan: Methodology, editing, and supervision.

**Conflict of Interest:** No potential conflict of interest was reported by the authors.

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

## References

Atılgan, H. (2008). Using generalizability theory to assess the score realibility of the special ability selection examinations for music education programmes in higher education. *International Journal of Research and Method Education*, *31*(1), 63-76. https://doi.org/10.1080/17437270801919925.

Atılgan, H., Kan, A. & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme.* (5. Baskı). Anı Yayıncılık.

Balbağ, M., Leblebicier, K., Karaer G., Sarıkahya E. & Erkan Ö. (2016). Türkiye'de fen eğitimi ve öğretimi sorunları. *Eğitim ve Öğretim Araştırmaları Dergisi, 5*(3), 1-12. http://www.jret.org/FileUpload/ks281142/File/02.m._zafer_balbag.pdf

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* ÖSYM

Bernardin, H. J. & Villanova, P. (2005). Research streams in rater self-efficacy. *Group and Organizational Management*, 30, 61-88. https://doi.org/10.1177/1059601104267675

Biemer, L. (1993). Trends-social studies /authentic assessment. *Educational Leadership*, *50* (8). https://www.ascd.org/el/articles/-authentic-assessment

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci başarıları* (Yayınlanmamış yüksek lisans tezi). Hacettepe Üniversitesi.

EARGED (2010). *PISA 2009 projesi, ulusal ön raporu*. 15 Mart 2011 tarihinde http://earged.meb.gov.tr/pdf/pisa2009rapor.pdf adresinden erişilmiştir.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercises Science, 5*(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2

Güler, N. (2013). *Eğitimde ölçme ve değerlendirme* (5. Baskı). Pegem Akademi.

Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment*, 7, 16-28. https://files.eric.ed.gov/fulltext/EJ1062689.pdf

Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi*. Pars Matbaacılık

Khodi, A. (2021). The affectability of writing assessment scores: A G-theory analysis of rater, task and scoring method contribution. *Language Testing in Asia 11*, Article 30 https://doi.org/10.1186/s40468-021-00134-5

Konak, Ö. A. (2010). Eğitim ve öğretim etkinlikleri üzerine. *Cito Eğitim: Kuram ve Uygulama Dergisi,* 10, 4-5.

Kutlu, Ö. (2006). Üst düzey zihinsel süreçleri belirleme yolları: Yeni durum belirleme yaklaşımları. *Çağdaş Eğitim Dergisi, 31(*335), 15-21. https://search.trdizin.gov.tr/tr/yayin/detay/74516/

Lee, Y. W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. ETS. http://www.ets.org/Media/Research/pdf/RM-04-07.pdf

Mcbee, M., & Barnes, L. (1998), The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11*(2), 179-194. https://doi.org/10.1207/s15324818ame1102_4

MEB (2017). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) projesi.* 1 Eylül 2022 tarihinde http://abide.meb.gov.tr/proje-hakkinda.asp adresinden erişilmiştir.

Mushquash, C., & O'Connor, B.P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods 38*, 542–547 https://doi.org/10.3758/BF03192810

Nalbantoğlu, F. & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 509-518. http://www.efdergi.hacettepe.edu.tr/shw_artcl-718.html

OECD (2012). *PISA 2009 technical report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

OECD (2017), OECD (2017), *PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving*, PISA, OECD Publishing http://dx.doi.org/10.1787/9789264281820-en

ÖSYM (2013). *Açık uçlu sorularla deneme sınavı: Soru/cevap kitapçığının yayımlanması* www.osym.gov.tr/belge/1-19413/acik-uclu-sorularla-deneme-sinavi-sorucevap-kitapcigini-.html adresinden erişim sağlanmıştır.

ÖSYM. (2017). *Açık uçlu sorular hakkında bilgilendirme ve açık uçlu soru örnekleri.* https://www.osym.gov.tr/TR,12909/2017-lisans-yerlestirme-sinavlari-2017-lys-acik-uclu-sorular-hakkinda-bilgilendirme-ve-acik-uclu-soru-ornekleri-05012017.html adresinden erişim sağlanmıştır.

Özçelik, D. A. (2010). *Ölçme ve değerlendirme*. Pegem Akademi.

Polat, M. & Turhan, N. (2021) Applying generalizability theory in language testing: Comparing nested and crossed scoring designs in the assessment of speaking skills, *International Journal of Curriculum and Instruction*,*13*(3), 3344–3358. https://ijci.globets.org/index.php/IJCI/article/view/825/409

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94* (1), 31-37. https://doi.org/10.5951/MT.94.1.0031

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing 22*(1) 1-30. https://doi.org/10.1191/0265532205lt295oa

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956–970 https://doi.org/10.1037/0021-9010.85.6.956

Sharma, F. & Weathers, D. (2003). Assessing generalizability of scales used in cross-national research. *International Journal of Research in Marketing*, 20, 287-295. http://dx.doi.org/10.1016/S0167-8116(03)00038-7

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications

Smith, Teresa A. (1997 March 24-28). *The Generalizability of Scoring TIMSS Open-Ended Items. (Report).* Annual Meeting of the American Educational Research Association, Chicago, USA

Turgut, F. M. (1992) *Eğitimde ölçme ve değerlendirme metotları. (*9. Baskı). Saydam Matbaacılık.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Wexley, K. N. & Youtz, M. A. (1985). Rater beliefs about others: Their effect on rating errors and rater accuracy. *Journal of Occupational Psychology*, 58, 265-275. https://psycnet.apa.org/doi/10.1111/j.2044-8325.1985.tb00200.x

Zorba, İ. (2020). *Personel alımında kullanılan bir yazılı sınav sonucunun genellenebilirlik kuramındaki farklı desenlerle karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Ankara Üniversitesi.