The Use of Open-ended Questions in Large-Scale Tests for Selection: Generalizability and Dependability

Hakan Atılgan [i]
Ege University

Elif Kübra Demir [ii]
Ege University

Tuncay Ogretmen [iii]
Ege University

Tahsin Oğuz Başokçu [iv]
Ege University

Abstract

It has become a critical question what the reliability level would be when open-ended questions are used in large-scale selection tests. One of the aims of the present study is to determine what the reliability would be in the event that the answers given by test-takers are scored by experts when open-ended short answer questions are used in large-scale selection tests. On the other hand, another aim of the study is to reveal how reliability changes upon changing the number of items and raters and what the required number of items and raters is to reach a sufficient degree of reliability. The study group consisted of 443 8th grade students from three secondary schools located in three different towns of the city of Izmir. These students were given a test including 20 open-ended short answer questions which was developed within the scope of the study. Students' answers were rated by four experienced teachers independently of one another. In the analyses, G theory's fully crossed two-facet design p x i x r with students (p), items (I) and raters (r). The analyses found $E\rho^2 = 0{,}890$ and $\Phi=0{,}855$ and it was concluded that well-educated raters in rating open-ended short answer questions can achieve consistent scoring at an adequate level.

Keywords : Large-Scale Tests, Open-Ended Question, Generalizability Theory, Rater Reliability, Generalizability, Dependability

----

[i] **Hakan Atılgan,** Prof. Dr., Faculty of Education, Ege University

[ii] **Elif Kübra Demir,** Research Assist Dr., Faculty of Education, Ege University, ORCID: 0000-0002-3219-1644

Correspondence: elif.kubra.demir@ege.edu.tr

[iii] **Tuncay Ogretmen,** Prof. Dr., Faculty of Education, Ege University

[iv] **Tahsin Oğuz Başokçu,** Assoc. Prof. Dr., Faculty of Education, Ege University

INTRODUCTION

Similar to many countries, selection exams are used in transition among education levels, especially for the university entrance, in Turkey as well. Such exams are used in transition among levels to select the right number of students for the seats available since the number of available seats is limited in the face of the high number of candidates. As they are tests taken by large masses and carried out to select students, these tests are called Large-Scale Selection Exams. In large-scale exams, different types of multiple-choice questions can be used as well as different types of open-ended questions. An important advantage of multiple choice questions in large-scale tests is that questions can be written in a wide part of the cognitive domain (Tekin, 1993; Turgut & Baykul, 2010; Atılgan, Kan, & Aydın, 2017; Miller, Linn, & Gronlund, 2009). It is relatively easy to write questions in the knowledge, comprehension and application levels of the cognitive domain while experts and experienced testers can write questions in the analysis and evaluation levels as well. Multiple-choice tests can be rated in a completely objective way (Tekin, 1993; Turgut & Baykul, 2010) and it is also possible to complete rating in a short time and announce the scores more rapidly. Large-scale exams are tests participated by masses. In such exams, speed is an important advantage in rating the test and announcing the results.

Besides the abovementioned advantages, multiple choice questions in large-scale exams also has some weaknesses. As stated above, although they can measure a large part of the cognitive domain, multiple choice questions cannot measure the synthesis level, which is a significant part of the cognitive domain. This is because the answerer cannot produce his/her own answer since the correct answer is provided among the options. As this prevents the individual to produce an authentic and creative answer, to transfer his/her opinions and organize his/her ideas, it is not possible to measure the behaviors at synthesis level (Rodriguez, 2016; Nitko & Brookhart, 2011; Berberoğlu, Milli Eğitim Bakanlığı Seviye Belirleme Sınavı (SBS) uygulamalarının değerlendirilmesi, 2009). Large-scale selection tests conducted with multiple choice questions are criticized for creating a group of graduates who have difficulty in expressing themselves and do not possess developed problem solving skills (Gür, Çelik, & Coşkun, 2013; Eğitim Reformu Girişimi, 2013; Dünya Bankası, 2011). Since large-scale exams are ranking tests in the general sense, ranking may change depending on chance success. This is unethical and differences caused by chance success may affect validity and reliability (Baykul, 2000; Mehrens & Lehmann, 1991) as well as causing wrong evaluations. It is also discussed that large-scale selection exams implemented with multiple choice questions have some negative side effects. These side effects can be broadly summarized as the exam becoming an instrument rather than a goal, the education system adapting itself in accordance with selection exams, candidates turning to private tutor and/or courses, increase in non-attendance at school and coming out of a group that cannot socialize (Yükseköğretim Kurulu, 2007; Türk Eğitim Derneği, 2010; Elçi, Süzme, Yıldız, Canpolat, & Çelik, 2016; Berberoğlu, Demirtaşlı, İşgüzel, Arıkan, & Özgen, 2010) . For open-ended questions, the answer is framed in the mind of the answerer, which is later produced by the answerer himself. Typically, open-ended questions are different from multiple choice questions in that they do not provide choices and the answer is produced by the answerer upon organizing information and expressing it in his/her own words/sentences. In other words, the main difference is the production of an answer instead of choosing from the given choices. Many resources classify open-ended questions in two groups based on their limitedness or freeness as a) short answer questions and b) essays (Tekin, 1993; Turgut & Baykul, 2010; Atılgan, Kan, & Aydın, 2017; Miller, Linn, & Gronlund, 2009).

Depending on the freeness of the answers to be given, open-ended questions are grouped as (a) restricted response and (b) extended response (Nitko & Brookhart, 2006; Kubiszyn & Borich, 2015). ÖSYM (Student Selection and Placement Center) (2015), on the other hand, added short answer questions to this classification considering the length of the answer and questions are categorized in three groups as (a) short answer questions, (b) restricted response questions and (c) extended response questions. Since the answer is not given within the question in open-ended question types the answerer is required to produce his/her own authentic answer. Therefore, open-ended questions can measure behaviors at the synthesis level and higher order mental skills by their nature (Turgut & Baykul, 2010; Kubiszyn & Borich, 2015). In addition, open-ended questions are able to measure composition skills, self expression through writing and the ability to use language in writing.

Since the correct answer to the question is not presented in the question as a choice, it is not possible for the responders to find the answer by chance. The absence of chance success is in important advantage for selection exams as it prevents the ranking to change depending on chance in tests. Also, since there are no choices for the answers, candidates who do not know the correct answer cannot find it by proceeding with choices or they cannot use association to find the answer, which makes open-ended questions advantageous in large-scale selection exams. In addition, it is easier to write questions that can measure higher order behaviors in comparison with multiple-choice test items. In this respect, open-ended questions are highly usable. Open-ended questions and tests consisting of this type of questions are cost-efficient in terms of printing and implementation.

Answering behavior in open-ended question includes reading the questions, framing the answer in mind and writing this answer by the individual. While varying based on the abovementioned question types, this process increases the responding time per item. Thus, the number of questions that could possibly be asked in a reasonable exam duration may be reduced, which could narrow the content to be measured. Such a downsizing in content decreases content validity while decreasing the number of questions affects reliability negatively in terms of precision (Baykul, 2000; Atılgan, Kan, & Aydın, 2017). Depending on the question types mentioned above, open-ended questions require much or less amount of writing activity. While the aim is to examine an individual's some other knowledge and skills, this type of questions may lead the answerer to fail to express these skills in writing although he/she does have them. In such situations where written expression skills are not intended for measurement, but some other skills are expected to be measured, wiring activity may interfere with the measurement results of writing skills and affect validity adversely (Turgut & Baykul, 2010). Rating open-ended questions is challenging and time-consuming. The efficient use of open-ended questions in large-scale selection exams to measure higher order behaviors can be effective in overcoming the difficulties or impossibilities faced in measuring higher order skills with multiple choice questions. However, particularly in large-scale tests, an important limitation is that rating the answers done by experts is not economical in terms of time and effort and the reliability of scores causes concerns.

As also stated above, using tests including only multiple-choice questions in large-scale selection examinations has significant limitations and some negative effect. Therefore, in order to eliminate such limitations in large-scale selection exams, it is necessary to include open-ended questions. However, when open-ended questions are used in large-scale tests, what the reliability would be or how would it be affected by human scoring comes up as a critical question. Based on these reasons, the present study aimed to determine the reliability levels and the most efficient combinations of rater and item numbers to reach the sufficient level of reliability in the case that open-ended questions (short answer) are used instead of multiple-choice test items and scored by experts in large-scale selection exams.

METHOD

The present study focuses on how reliability is affected when open-ended short answer questions in large-scale exams are scored by multiple raters and what the optimal number of raters/items should be to reach the required reliability. Since the study is multi-faceted, the best way of determining reliability can be obtained by G theory. Therefore, Transition from Primary to Secondary Education (henceforth referred to as TEOG) exams, which are applied on 8th graders in Turkey, were studied. In the event that open-ended short answer questions are used in these exams, instead of such subjects as Mathematics and Science, in which answers are more concrete, Turkish was included in the study scope as it is considered to possibly cause lower rater consistency by its nature.

Study Group

The study group consisted of 443 8th graders among the students who took the TEOG exam. The 443 students included in the study group were selected from three towns of Izmir (Karşıyaka, Bayraklı and Bornova) using cluster sampling method. A total of three schools (one from each town) were selected using cluster sampling again and all students attending the 8th grade at these schools

were included within the scope of the study. For the G theory analyses used in the study, this sample size can be accepted as large enough (Atılgan, 2013).

Data Collection Instrument

In the event that open-ended short answer questions are used TEOG exams, instead of a subject, in which answers are more concrete, Turkish was included in the study scope as it is expected to possibly cause lower rater consistency by its nature. The Turkish section of TEOG includes 20 multiple choice questions with four choice options (Ministry of Education (MoE), 2014). Items included in the Turkish section of the TEOG exam held in November 2014 were assessed by three Turkish language teachers, 2 measurement and assessment experts and a curriculum development expert. Teachers were trained on test development. Considering the properties of the given items and their distribution of subject areas, 20 open-ended short answer questions were written by this group of experts. The questions were intended to measure higher order behaviors of the cognitive domain in comparison with the questions in TEOG.

The test which was developed in the 15 days following November 2014 TEOG and included open-ended short answer questions were applied to 443 8th grade students who took TEOG. In order to avoid any changes in the measured student characteristics, the time gap between the November 2014 TEOG and the implementation of the open-ended short answer test used in the study was not allowed to exceed 15 days. The open-ended short answer test was given to students in their own classrooms and by their own teachers paying attention to ensuring student motivation.

The answers given to the open-ended short answer questions by 443 students in the study group were scored by a selected group of four experienced Turkish language teachers. Before scoring, the four teachers to perform the scoring informed about how to do the scoring in a meeting. Each teacher rated the papers of all 443 students. During the scoring period, it was ensured that the raters had no communication with one another and performed their scorings independently. The teachers were asked to score each answer between 0 and 3 based on their accuracy and authenticity and an answer key was used as well.

Data Analysis Method

When open-ended questions are asked, and the answers are scored by multiple raters in large-scale selection exams, the error source for reliability comes up both as the items and the raters. Generalizability (G) Theory, which is used in determining a single reliability coefficient in such cases based on multiple error sources, was developed by Cronbach et al. (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001). The data obtained for the measurement situation in which four raters scored all the answers given to 20 open-ended short answer questions by all the 443 students were analyzed using the Generalizability (G) and Decision (D) studies of G theory. In the analysis, fully crossed two-facet p x i x r design of G theory was used with students (p), items (i) and raters (r) (Shavelson & Webb, 1991; Brennan, 2001).

FINDINGS

Generalizability *Analyses*

Scorings of the four raters performed on all the answers given to 20 open-ended short answer questions by all the 443 students were analyzed using the Generalizability (G) and Decision (D) studies of G theory. In order to find an answer to the study problem fully crossed two-facet p x i x r design of G theory was used in the analyses with students (p), items (i) and raters (r) (Shavelson & Webb, 1991; Brennan, 2001).

G Study

As a result of the G study analyses performed using the single variable fully crossed p *x* r *x* I model of G theory, variances for facets and the share of each facet within the total variance were calculated.

The values obtained for the main effect variances $\left(\sigma_p^2,\ \sigma_i^2,\ \sigma_r^2\right)$ and interaction effect variances $\left(\sigma_{pi}^2,\ \sigma_{pr}^2,\ \sigma_{ir}^2,\ \sigma_{pir}^2\right)$ of the facets and the percentages of these variances within the total variance are presented in Table 1.

Table 1. Estimated Variances and their Percentages in the Total Variance

Source of Variation	Variance	Percentage
Student (*p*)	0,379	20,340
Item (*i*)	0,283	15,206
Rater (*r*)	0,010	0,537
Student x Item (*pi*)	0,832	44,662
Student x Rater (*pr*)	0,006	0,319
Item x Rater (*ir*)	0,045	2,394
Student x Item x Rater (*Pir,ₑ*)	0,308	16,541
Total	1,863	100,000

As seen in Table 1, student (*p*) main effect variance component was estimated as $\left(\sigma_p^2\right)$ 0,379 with Generalizability (G) analysis. This estimated student (*p*) main effect variance component explains 20,340% of the total variance. Student main effect is a population score variance and shows to what extent students differentiate from each other in terms of their measured abilities (Shavelson & Webb, 1991; Brennan, 2001; Atılgan, 2008). The fact that this variance component estimated for students is the second largest variance within the total variance, as an expected result, is an indicator that differences of the 443-student sample in terms of their abilities measured can be revealed at this scale.

Item main effect variance has the second largest variance among main effects with $\left(\sigma_i^2\right)$ 0,283 and its share in the total variance was calculated as 15,206%. Shavelson and Webb (1991) and Brennan (2001) state that the main effect variance estimated for item facet show the changeability of item difficulties. In other words, a large item main effect variance component reveals that some questions are easier or more difficult than others while a small one shows that items are close to one another in terms of difficulty. In the present study, the proportional greatness of item main effect variance (15,206%) in the total variance shows that in the test consisting of 20 open-ended items, difficulty levels of these 20 items differ from each other and the questions vary in terms of easiness and difficulty.

As seen in Table 1, in the analysis, estimated variance for rater main effect was found as $\left(\sigma_r^2\right)$ 0,00006. The share of this rater facet variance within the total variance is 0,065%. This percentage is quite small compared to the total variance.

Shavelson and Webb (1991) and Brennan (2001) report that this variance estimated for rater facet main effect is an indicator that a certain rater acts more generously or more strict in the scores they give for all students than the other raters. The fact that the main effect variance estimated for rater facet is close to zero and has a quite small percentage within the total variance shows that the difference among the scores given by the four independent raters to all students is quite small and that the raters are consistent with each other.

Student and item variance as an interaction effect variance was obtained as $\left(\sigma_{pi}^2\right)$ 0,0832. Its share within the total variance was found as 44,662%. This variance is the greatest of all other variances and constitutes the highest percentage among variances. *Pi* variance as an interaction effect variance shows whether a certain student's relative state changes from one item to another (Shavelson

& Webb, 1991; Brennan, 2001; Atılgan, 2008). The fact that *pi* variance as an interaction effect variance is the greatest one in the present study reveals that variaance in 443 students' relative state from one item to another among the total 20 items is high.

Another interaction effect variance, student *and* rater interaction effect variance, was calculated as (σ_{pr}^2) 0,006 with a percentage of 0,319% in the total variance. This interaction effect variance can be said to be quite small and have a low share in the total variance. Student and rater interaction effect shows whether a certain rater scored a certain student more generously or more strictly than the other raters (Shavelson & Webb, 1991; Brennan, 2001; Atılgan, 2008). That this variance is close to zero and its percentage in the total variance is small indicates that the scores given by a certain rater to a certain student are rather consistent with the other raters. In other words, it can be suggested that in the scoring of any one of the four raters for any one of the 443 students, differentiation in terms of strictness/generosity is quite small compared to the other raters and students.

Item and rater interaction effect variance was found as (σ_{ir}^2) 0,045. The share of this variance in the total variance was calculated as 2,394%. The *ir* interaction effect variance calculated by G study shows whether raters score consistently from one item to another (Shavelson & Webb, 1991; Brennan, 2001; Atılgan, 2008). The fact that the interaction effect variance is relatively small shows in the scorings performed by four raters on the 20-item open-ended test for 443 students, differentiation among scorings is relatively small from one item to another. That is, it can be said that scorings performed by each of the four raters on each of the 20 items are highly consistent with each other.

In the G study, student, item and rater variance, also called residual variance, was found as $(\sigma_{pir,e}^2)$ 0,308. The percentage of the residual variance in the total variance is 16,541%. Residual variance is caused by systematic or random errors and or those that cannot be explained with the available data and is expected to be small (Shavelson & Webb, 1991; Brennan, 2001; Atılgan, 2008). In the present study, the fact that the residual variance is relatively large shows the quantity of systematic and random errors and/or the variance that cannot be explained with the main and interaction effect variances in the measurement performed by 4 raters on 20 items for 443 students.

G theory was used to determine the reliability of the scorings performed by four raters on 20 open-ended items for 443 students. G theory considers that there are two types of decision making, relative and absolute evaluation, for determining reliability in education and psychology. Therefore, two different coefficients of reliability are calculated by G theory; generalizability (G) coefficient for relative evaluations and (Phi) for absolute evaluations   (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001).

Relative error variance is represented by $\sigma^2(\delta)$ and is calculated as in Equation 1 depending on *pi*, *pr*, *pir,$_e$* variances. G coefficient, which is used for relative evaluations and represented by Eρ^2, is defined with Equation 2 as the ratio of the population score variance of individuals symbolized with $\sigma^2(\tau)$ to the sum of the population score variance of individuals and relative error variance$[\sigma^2(\delta)]$ (Brennan, 2001; Atılgan, 2005).

$$\sigma^2(\delta) = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r} \qquad \text{Equation 1}$$

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \qquad \text{Equation 2}$$

Phi coefficient used for absolute evaluations and represented by Φ; is defined as Equation 4 as the ratio of the population score variance of the people represented by $\sigma^2(p)$ to the total absolute error variance obtained from Equation 3 based on this variance and the *i*, *r*, *pi*, *pr*, *ir*, *pir,$_e$* variances represented by $\sigma^2(\Delta)$. (Brennan, 2001; Atılgan, 2005).

$$\sigma^2(\Delta) = \frac{\sigma_i^2}{n_i} + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{ir}^2}{n_i n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r} \qquad \text{Equation 3}$$

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \qquad \text{Equation 4}$$

In the present study, G coefficient $(E\rho^2)$ concerning the scorings performed by four raters on 20 open-ended items for 443 students was found as 0,89. On the other hand, Phi coefficient (Φ) for the scorings performed by four raters on 20 open-ended items for 443 students was calculated as 0,855
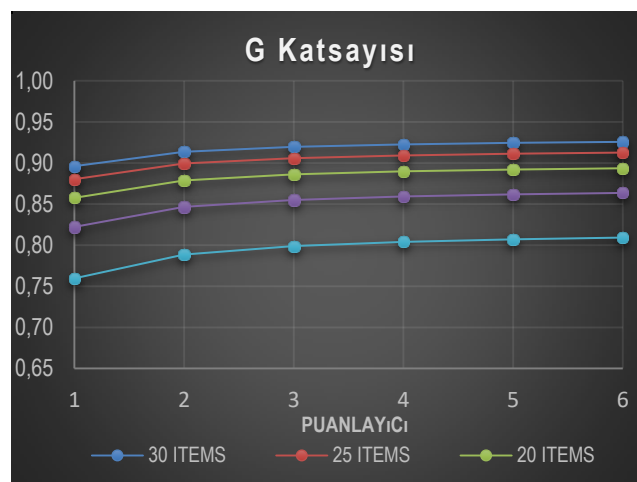
K Study

In the K study analyses performed using the single variable two facet crossed pxrxi design of G theory, G $(E\rho^2)$ coefficients were calculated for the situations when the numbers of items and raters of the test, which consisted of 20 open-ended items and was scored by 4 raters, were increased or decreased by 5 and 1 respectively. G $(E\rho^2)$ coefficients obtained from the K study performed with alternative numbers of items and raters are presented in Table 2 below.

Table 2. $E\rho^2$ Coefficients obtained from the K Study*

Rater	30 Item $E\rho^2$	25 Item $E\rho^2$	*20 Item* $E\rho^2$	15 Item $E\rho^2$	10 Item $E\rho^2$
6	0,926	0,913	0,894	0,864	0,809
5	0,924	0,911	0,892	0,862	0,807
4	0,923	0,909	*0,890*	0,859	0,804
3	0,920	0,906	0,886	0,855	0,799
2	0,914	0,899	0,879	0,846	0,789
1	0,896	0,880	0,858	0,822	0,760

*Figures in bold and italic show the number of items and raters

Alternative K studies carried out aim to specify the optimal number of items and raters to reach a sufficient level of reliability by examining the changes in $E\rho^2$ coefficients depending on the changes in the number of raters and items. As seen in Table 2 and Figure 1, $E\rho^2$ coefficients may increase more when the number of items is increased rather than that of the raters. To illustrate, when the number of raters is reduced from 4 to 2 and the number items is raised from 20 to 25, $E\rho^2$ increases from 0,890 to 0,899. That is, in the present implementation, the $E\rho^2$ coefficient reached with 4 raters and 20 open-ended items and the $E\rho^2$ coefficient to be reached with 2 raters and 25 items are approximately at the same level.

Figure 1. Changes of $E\rho^2$ coefficient obtained from the K Study

In the K study analyses performed using the single variable two facet crossed pxrxi design of G theory, Phi (Φ) coefficients were calculated for the situations when the numbers of items and raters of the test, which consisted of 20 open-ended items and was scored by 4 raters, were increased or decreased by 5 and 1 respectively. Phi (Φ) coefficients obtained from the K study performed with alternative numbers of items and raters are presented in Table 3 below.

Table 3. Phi (Φ) Coefficients obtained from the K study *

Rater	30 Item Φ	25 Item Φ	*20 Item* Φ	15 Item Φ	10 Item Φ
6	0,901	0,884	0,861	0,824	0,759
5	0,899	0,882	0,858	0,822	0,757
4	0,896	0,879	*0,855*	0,818	0,753
3	0,891	0,874	0,850	0,812	0,747
2	0,881	0,864	0,839	0,801	0,734
1	0,854	0,835	0,809	0,769	0,700

*Figures in bold and italic show the number of items and raters

Alternative K studies carried out aim to specify the optimal number of items and raters to reach a sufficient level of reliability by examining the changes in Phi (Φ) coefficients depending on the changes in the number of raters and items. Figure 2 shows the graph for an easier observation of the change in Phi coefficients obtained from the K study results carried out with alternative numbers of items and raters.
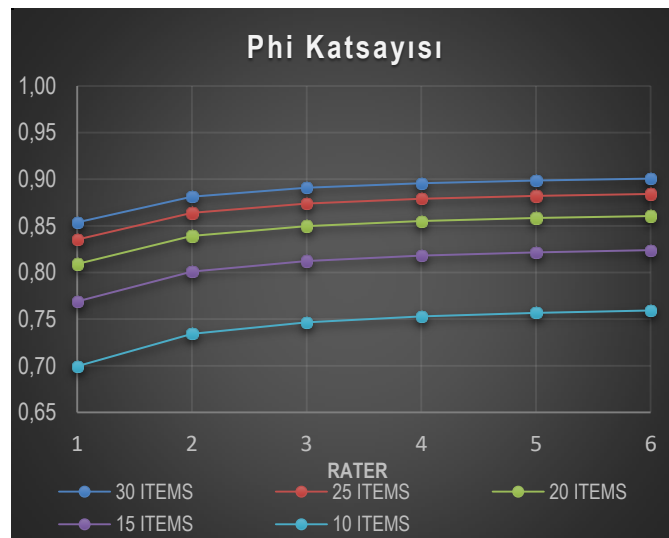
Figure 2. Changes of Phi (Φ) Coefficients obtained from K Study

As seen in Table 3 and Figure 2, similar to G coefficient, the increase occurring in Phi coefficient may be higher when the number of items is increased rather than the number of raters. For instance, when the number of raters is reduced from 4 to 2, but the item number is increased to 25 from 20, Phi coefficient rises to 0,864 from 0,855. In other words, in the present implementation, the Phi coefficient obtained with 4 raters and 20 open-ended items and the Phi coefficient to be reached with 2 raters and 25 items are approximately at the same level.

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

In the fully crossed two facet design G study conducted in the event that a 20- open-ended – short-answer item test applied on 443 students are scored by four raters independently of one another item fact accounts for %15,206 while rater facet constitutes 0,065% of the total variance in the present study. This finding shows that items vary in terms of difficulty whereas raters do not differ from one

another in their scorings and performed consistent scorings. On the other hand, the percentages of interaction effect variances within the total variance were obtained as $\sigma_{pi}^2 = \%44{,}62$; $\sigma_{pir,e}^2 = \%16{,}541$; $\sigma_{ir}^2 = \%2{,}394$; $\sigma_{pr}^2 = \%0{,}319$ from the greatest to the smallest. These results reveal that students' relative states change from one item to another, the variance which cannot be explained by main and interaction effect is high and/or has systematic and random errors, and scorings of the raters on each of the items are consistent with one another. In addition, greatness of student and item interaction effect variance and residual variance decreases reliability. For the measurement state including 20 items and four raters, G and Phi coefficients were found as 0,890 and 0,855 respectively in the study.

In the K study analyses performed using the single variable two facet crossed pxrxi design of G theory, when the numbers of items and raters of the test which consisted of 20 open-ended items and was scored by 4 raters were increased or decreased by 5 and 1 respectively, it was seen that increasing the number of items rather than that of the raters is more effective in terms of G and Phi coefficients.

G ($\boldsymbol{E\rho^2}$) coefficient is used for relative measurements, and Φ coefficient for relative measurements (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001). Since large-scale selection exams are the focus of the present study and large-scale exams are relative measurements, $\boldsymbol{E\rho^2}$ coefficient is more significant.

In this respect, it can be concluded that two raters with good expertise and scoring can perform consistent scorings at a sufficient level in large-scale selection exams consisting of open-ended short answer questions. On the other hand, in the event that an enough number of well-designed (considering content validity) open-ended short answer questions are used it is possible to measure higher order cognitive skills which cannot be measured by multiple choice questions and to attain the required level of reliability as well. Since the training of raters who scored open-ended questions on the scoring tool and how to score increases their scoring consistency, such training may be recommended before scoring. Open-ended questions can be used instead of multiple-choice questions in cases where sufficient consistency between raters can be achieved in scoring and the reliability of measurement results can be achieved.

Acknowledge

REFERENCES

Arslan, M. (2004). Eğitim Sistemimizin Kapanmayan Yarası-Yükseköğretime Geçiş. *Sosyal Bilimler Enstitüsü Dergisi, 16*37-51.

Atılgan, H. (2005). Genellenebilirlik Kuramı ve Puanlayıcılar Arası Güvenirlik için Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama, 4* (7), 95-108.

Atılgan, H. (2008). Using generalizability theory to assess the score reliability of the special ability selection examinations for music education programs in higher education. *International Journal of Research & Method in Education, 31*(1), 63-76.

Atılgan, H. (2013). Sample size for estimation of G and Phi coefficients in generalizability theory. *Eurasian Journal of Educational Research, 51*, 215-228.

Atılgan, H., Kan, A., & Aydın, B. (2017). *Eğitimde Ölçme ve Değerlendirme*, (Edt. Hakan Atılgan). Ankara: Anı Yayıncılık.

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items.* Princeton, NJ: ETS.

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması.* Ankara: ÖSYM.

Berberoğlu, G. (2009). Milli Eğitim Bakanlığı Seviye Belirleme Sınavı (SBS) Uygulamalarının Değerlendirilmesi. *Cito Eğitim: Kuram ve Uygulama*, 2, 9-24.

Berberoğlu, G., Demirtaşlı, N., İşgüzel, Ç., Arıkan, S., & Özgen, T. (2010). Okul Dışı Etmenlerin Okul Başarısı ile İlişkisi. *Cito Eğitim: Kuram ve Uygulama*, 7, 27-38.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Burrows, S., Gurevyc, I., & Stei, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25-60.

Burstein, J., Leacock, C., & Swartz, R. (2001). *Automated evaluation of essay and short answers.* Princeton, NJ: ETS Technologies, Inc. A Subsidiary of Educational Testing Service.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Demirtaşlı, N. Ç. (2010). Açık uçlu soru formatı ve öğrenci izleme sistemi (ÖİS) akademik gelişimi izleme (AGİD) modülündeki kullanımı. *Cito Eğitim: Kuram ve Eğitim*, s. 21-28.

Dünya Bankası. (2011). *Türkiye'de Temel Eğitimde kalite ve eşitliğin geliştirilmesi: zorluklar ve seçenekler.* Rapor No:54131-TR. Ankara: Dünya Bankası İnsani Kalkınma Departmanı Avrupa ve Orta Asya Bölgesi.

Eğitim Reformu Girişimi, (2013). *Yeni Ortaöğretime Geçiş Sistemi Üzerine Değerlendirme*. İstanbul: Eğitim Reformu Girişimi.

Elçi, Y., Süzme, P. S., Yıldız, R., Canpolat, Y., & Çelik, O. (2016). *Ortaöğretimi izleme ve değerlendirme raporu* (Ed: Hacı Ali Okur). Ankara: Milli Eğitim Bakanlığı Ortaöğretim Genel Müdürlüğü.

Gomma, W., & Fahmy, A. (2014). Arabic Short Answer Scoring with Effective Feedback for Students. *International Journal of Computer Applications*, (86), 35-41.

Gür, B. S., Çelik, Z., & Coşkun, İ. (2013). *Türkiye'de Ortaöğretimin Geleceği: Hiyerarşi mi eşitlik mi?* Sayı 69. Ankara: SETA Analiz.

Güven, İ. (2010). *Türk Eğitim Tarihi*. Ankara: Naturel.

Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Hights, MA: Ally & Bacon.

Jang, E. S., Kang, S. S., Noh, E. H., Kim, M. H., Sunk, K. H., & Seong, T. J. (2014). KASS: Korean Automatic Scoring System for Short-answer Questions. 6th International Conference on Computer Supported Education, (s. 226-230).

Kubiszyn, T., & Borich, G. D. (2015). *Educational testing and measurement: classroom application and practice*. Hoboken, NJ: John Wiley & Sons.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology.* Belmont, CA: Wadsworth.

Miller, D. M., Linn, R. L., & Gronlund, N. E. (2009). *Measuremet assessment in teaching*. New Jersey: Pearson Education Inc.

Milli Eğitim Bakanlığı, (2014). *8.Sınıf I.Dönem Ortak Sınavı Soruları ve Cevap Anahtarı.* Milli Eğitim Bakanlığı Web Sitesi: http://www.meb.gov.tr/sinavlar/dokumanlar/2014/soru/8SinifOrtakSinavlar_1_Donem/Turkce/TURKCE_A.zip adresinden alındı

Milli Eğitim Bakanlığı. (2016a). *Uluslararası öğrenci değerlendirme programı PISA 2015 ulusal raporu.* Ankara: Milli Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı. (2016b). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. Ankara: Milli Eğitim Bakanliği Ölçme, Değerlendirme ve Sinav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı Teftiş Kurulu Başkanlığı. (2010). *Ortaöğretime Geçiş Sisteminde SBS ve Yeni Bir Model*. Ankara: Milli Eğitim Bakanlığı.

Nitko, A. J., & Brookhart, S. M. (2006). *Educational assessment of students* (5th edition). Upper Saddle River, NJ,: Prentice-Hall, Inc.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of student*. Boston, MA: Pearson Education.

Nitko, A. J., & Brookhart, S. M. (2016). *Öğrencilerin eğitisel değerlendirilmesi*. (B. Bıçak, M. Bahar ve S. Özel, Çev. Edt.). Ankara: Nobel Yayıncılık.

Nunnally , J., & Bernstein , I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ölçme, Seçme ve Yerleştirme Merkezi. (2017). *2017 Öğrenci Seçme ve Yerleştirme Sistemi* (ÖSYS) Klavuzu. www.osym.gov.tr: http://dokuman.osym.gov.tr/pdfdokuman/2017/OSYS/LYS/KILAVUZ_18042017.pdf adresinden alındı

ÖSYM. (2015). *Yazılı Sınav (Açık Uçlu Sorularla Sınav).* www.osym.gov.tr: http://www.osym.gov.tr/TR,721/yazili-sinav-acik-uclu-sorularla-sinav-04022015.html adresinden alındı

Polat, S., Özoğlu, M., Yıldız, R., & Canbolat, Y. (2013). *Ortaöğretim izleme ve değerlendirme raporu*. Ankara: Ortaöğretim Genel Müdürlüğü.

Rodriguez, M. C. (2016). *Sellected-response item development*. S. Lana, T. Haladyna, & M. Raymond içinde, Handbook of test development, 2th edition. New York : Taylor & Francis / Routledge.

Shavelson, R. J., & Webb, M. N. (1991). *Generalizability Theory Aprime*. California: Sage Publication.

Srihari, S., Collins, J., Srihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffler. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence*, 172, 300-324.

Steeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). *Pearson's Automated Scoring of Writing, Speaking and Mathematics*. UK: Pearson.

Sukkarieh, J., & Blacmore, J. (2009). *c-rater:automatic content scoring for short constructed responses.* Proceedings of the Twenty-Second International FLAIRS Conference, (s. 290-295).

Tekin, H. (1993). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Yargı Kitap ve Yayınevi.

Thordike, R. M., & Thordike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education.* (8th Edition). Boston: Pearson Education, Inc.

Traub, R. E. (1994). *Reliability for Social Sciences: Theory and Applications.* California: Sage Publications.

Turgut, M., & Baykul, Y. (2010). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Pegem Akademi.

Türk Eğitim Derneği. (2010). *Ortaöğretime ve Yükseköğretime Geçiş Sistemi Özet Raporu.* Ankara: Türk Eğitim Derneği.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6. 103-118.

Yükseköğretim Kurulu. (1999). *Yükseköğretime Giriş Sınavı: Geçmiş Yıllarla Karşılaştırma ve Değerlendirme*. Ankara: Yükseköğretim Kurulu.

Yükseköğretim Kurulu. (2007). *Türkiye'nin Yükseköğretim Stratejisi*. Ankara: Yükseköğretim Kurulu.