Eurasian Journal of Educational Research 80 (2019) 133-150

Eurasian Journal of Educational Research
www.ejer.com.tr

Reliability of Essay Ratings: A Study on Generalizability Theory

Hakan ATILGAN[1]

ARTICLE INFO

Article History:
Received: 2 Aug. 2018
Received in revised form: 8 Jan. 2019
Accepted: 8 Mar. 2019
DOI: 10.14689/ejer.2019.80.7

Keywords
Generalizability Theory, generalizability, reliability, essay rating, essay rater reliability, writing ratings

ABSTRACT

Purpose: This study intended to examine the generalizability and reliability of essay ratings within the scope of the generalizability (G) theory. Specifically, the effect of raters on the generalizability and reliability of students' essay ratings was examined. Furthermore, variations of the generalizability and reliability coefficients with respect to the number of raters and optimal number of raters for obtaining optimal reliability of the rating of the writing ability of a student, which is considered to be an implicit trait as a whole and in its sub-dimensions of wording/writing, paragraph construction, and title selection, were determined.

Research Methods: The student sample of the study comprised 443 students who were selected via random cluster sampling, and rater sample of this study comprised four Turkish teachers. All the essays written by the students in the sample were independently rated on a writing skill scale (WSS), which is an ordinal scale comprising 20 items, by four trained teachers. In this study, data analysis was performed using the multivariate $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ design of the G theory.

Finding: In the G studies that were performed, variances of the rater (*r*) as well as item and rater (*ixr*) were low in all sub-dimensions; however, variance of the object of measurement and rater (*pxr*) was relatively high. The presence of trained raters increased the reliability of the ratings.

Implications for Research and Practice: In the decision (D) study analyses of the original study conducted using four raters, the G and Phi coefficients for the combined measurement were observed to be .95 and .94, respectively. Further, the G and Phi coefficients were .91 and .90, respectively, for the alternative D studies that were conducted by two trained raters. Thus, rating of essays by two trained raters may be considered to be satisfactory.

[1] Ege University, TURKEY, E-mail: hakan.atilgan@ege.edu.tr, ORCID: https://orcid.org/0000-0002-5562-3446

Introduction

Different tools are used depending on the feature of education that is to be measured. One of these measurement tools is essay-type examinations, which are appropriate for measuring high-level skills, including writing, self-expression in a native or foreign language, problem solving, creative thinking, critical thinking and synthesis step behaviours (Atilgan, Kan & Aydin, 2017; Turgut & Baykul, 2010). Cohen, Swerdlik and Philipps (1996) also emphasised that essay-type examinations require organisation, planning and writing skills. Writing is a critical skill (Graham, Harris & Hebert, 2011); therefore, writing and writing-based essay-type examinations constitute a primary mechanism by which students can display their knowledge (Graham, 2006).

Furthermore, essay-type examinations are tests by which students are expected to display their academic content knowledge (Bereiter, 2003). Generally, a student writing an essay must gather his/her thoughts about a given subject, create an idea, and organise his/her thoughts. Essay-type examinations are more recognised compared to other types of examinations for measuring writing ability of a student (Atilgan, Kan & Aydin, 2017; Schoonen, 2005). From this viewpoint, essay-type examinations are considered to be essential measurement tools in the field of education. However, even though essay-type examinations exhibit various advantages while measuring writing ability of a student, it exhibits various disadvantages, such as the creation of errors, because of the complexity and versatility of essay-type examinations (Shavelson, Baxter & Gao, 1993).

Because there are differences between writing abilities of various students, students are not expected to achieve identical ratings in essay-type examinations. Furthermore, ratings will vary from one student to another, thereby reflecting differences between their writing abilities. However, a student's rating is affected by several extraneous factors. With respect to writing, which is a complex ability, these extraneous factors include several variance sources such as the task, type of task, rater, rating tool, essay topic, student's interest in the topic, essay type (such as descriptive, analytical, narrative or argumentative), time constraint, rating process, interaction, and other such factors (Schoonen R., 2005; Sudweeks, Reeve & Bradshaw, 2005). Moreover, changes in ratings that are obtained based on this variance are considered to be measurement errors.

Similar to all ratings, main objective of measurements in essay-type examinations is to accurately evaluate the measured feature of students (Kim, Schatschneider, Wanzek, Gatlin & Otaiba, 2017; Nitko & Brookhart, 2011; Nunnally & Bernstein, 1994). However, as mentioned previously, apart from a student's writing efficiency, measurement errors arising from the sources of variance, such as raters, tasks and measurement tools, also affect measurement results (Schoonen, 2012). Presence of errors from such sources of variance while measuring writing abilities complicates the determination of reliability (Bouwer, Beguin, Sanders & van den Berg, 2015).

Nitko and Brookhart (2011, p. 219) indicate that intra-rater reliability is low because of the nature of essay-type questions. In particular, rater is the source of variance that

affects the reliability of essay-type examinations. Because the same rater may rate the same essay differently at different times (Block, 1985; Cooper, 1984), the same essay may achieve inconsistent ratings when multiple raters rate it independently from each other (Baykul, 2000; Tugut, 1995). Furthermore, scoring reliability can be increased when the raters are provided with a high level of training (Weilgle, 1994; Weigle, 1998). However, raters may interpret the rating criteria differently and rate differently despite their high level of training (Gebril, 2009; Schoonen R., 2005; Swartz, et. al., 1999). Several studies have shown that raters differ in their implementation of the rating criteria in terms of rigidity and generosity (Atilgan, 2008; Cumming, Kantor & Powers, 2002; Eckes, 2008; Kan, 2007; Kondo-Brown, 2002).

Measurement errors that are caused by this differentiation among raters result in inconsistency in rating and decrease in reliability. Furthermore, determination of the accuracy of rating obtained via essay-type examinations depends on the measurement errors that arise from the sources of variance. Simultaneously, to minimise the interference of such errors with the measurement results, sources of these errors should be accurately understood; moreover, measurement conditions should be designed accordingly. The generalizability theory (G theory) is an appropriate methodology for designing measurement tools by determining the errors arising from multiple sources of variance.

Generalizability Theory

While determining reliability, the classical test theory considers only the errors that are obtained from a single source of variance such as items, raters and time (Crocker & Algina, 1986; Lord & Novick, 1968; Miller, Linn & Gronlund, 2009; Thordike, 1971). For example, in case of test–retest reliability, source of variance (error) is considered to be time, whereas, in case of Cronbach's alpha reliability coefficient, source of variance (error) is items. However, in some measurements, multiple sources of variance can exist. For example, in several multifaceted measurements, items that are rated using multiple raters, items and raters as well as their interactions are considered to be sources of potential variance. The G theory, which can simultaneously consider all the sources of potential variances and their interactions (Atilgan, 2008; Brennan, 2001a; Crocker & Algina, 1986; Cronbach, 1984; Nunnally & Bernstein, 1994; Shavelson & Webb, 1991), has been proposed by Cronbach et. al., (Cronbach, Rajaratnam & Gleser, 1963; Cronbach, Gleser, Nanda & Rajaratnam, 1972) as an expansion of the classical test theory for overcoming its limitations.

In a measurement scenario, a G study is conducted for determining the effects of error sources by analysing all error sources together and for defining the universe of admissible observation. The G theory can divide observed ratings into facets, interaction of facets and random errors. For example, the most prevalent G theory is a completely crossed design (p × i × r), where performances of the objects of measurement (p) are rated by multiple raters (r) using multiple items (Atilgan, 2008). In this design, p, i and r are referred to as facets. The p × i × r design of the G theory contains seven variances $\left(\sigma_p^2,\ \sigma_i^2,\ \sigma_r^2,\ \sigma_{pi}^2,\ \sigma_{pr}^2,\ \sigma_{ir}^2,\ \sigma_{pir,e}^2\right)$ comprising three main and four interaction effect variances (Atilgan, 2008; Brennan, 2001a; Shavelson & Webb,

1991). In the G study, these variances can be estimated using analytic variance techniques. Furthermore, relative error variance (δ) is defined, as presented in Equation 1, using the variances of interaction between estimated components of variance, including the objects of measurement and other facets.

$$\delta = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r} \quad (1)$$

Furthermore, absolute error variance (Δ), as presented in Equation 2, is defined using the main effects of facets (except for the objects of measurement) and interaction variances among all the facets.

$$\Delta = \frac{\sigma_i^2}{n_i} + \frac{\sigma_r^2}{n_r} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{ir}^2}{n_i n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r} \quad (2)$$

The generalizability coefficient ($E\rho^2$) is defined for performing relative measurements, as presented in Equation 3, using relative variance (δ). Furthermore, reliability (Phi) coefficient (Φ) is defined for performing absolute measurements, as presented in Equation 4, using absolute error variances (Δ) (Atilgan, 2008; Brennan, 2001a; Shavelson & Webb, 1991).

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \delta} \quad (3)$$

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} \quad (4)$$

The decision study (D) is conducted for determining the optimum conditions of facets, including the number of items and raters, using variances obtained from the G study for minimising the errors in a measurement design (Brennan, 2001a; Crocker & Algina, 1986; Shavelson & Webb, 1991). Furthermore, change in measurement error and reliability can be estimated by increasing or decreasing the number of each facet, such as item and rater, using the D study. Thus, measurement designs can be determined in which the conditions of facets may be considered to be optimal for achieving the desired level of reliability.

Several studies have been conducted based on G theory from the viewpoint of rating writing abilities and reliability of ratings. In some of these studies, rater and task (Kim, Schatschneider, Wanzek, Gatlin & Otaiba, 2017), rater and occasion (Sudweeks, Reeve & Bradshaw, 2005) and rater's years of experience (Dogan & Uluman, 2017) are examined as facets. In some studies related to the reliability of the writing ratings, certain traits, such as the topic of writing task, content or use of language, whether rating is analytic or holistic (Schoonen, 2005), whether rating guidance is used (Kan, 2007), the number of essay samples (Graham, Hebert, Sandbank & Harris, 2016), essay type (such as argumentative, narrative) (Bouwer, Beguin, Sanders & van den Ber, 2015) and different task types (Gebril, 2009), are considered to be the facets. Although several studies have determined the intra-rater reliability, only a few generalizability studies have studied the ratings using trained raters. Studies related to scoring and

generalizability of writing skills have mostly focused on writing skills in foreign languages, and the G theory analyses have been conducted using univariate patterns having sample widths of lower than 200. It is assumed that this study, which is conducted using the multivariate G theory pattern, will contribute to the literature with a large sample, where raters have been trained to rate writing skills in their native language.

Herein, the generalizability and reliability of the essay ratings, which measure writing abilities of the objects of measurement in their native Turkish language, have been examined in the context of multivariate G theory. In this context, the effects of raters who have been trained on the subject of rating are considered to be effective with respect to generalizability and reliability of essay ratings. This study has attempted to denote the manner in which the coefficients of generalizability and reliability change according to the number of raters while rating writing ability, which is an implicit trait, as a whole and in its sub-dimensions of title selection, paragraph construction and wording/writing along with a suitable number of raters for ensuring optimal reliability. Thus, this study intended to broaden our knowledge related to assessment of essay writing skills and to create a reference for obtaining a sufficiently reliable rating of essays.

Method

Research Design

The present study aimed to investigate generalizability and reliability of the essay ratings. The following sections describe the research sample, data collection procedure, tool and research data, and data analysis.

Research Sample

Atilgan (2013) indicates that a sample size of 400 is sufficient for performing an accurate and reliable estimation of the G and Phi coefficients. Therefore, size of the student sample of the study is targeted to be greater than 400. Therefore, three districts, namely Bayrakli, Bornova and Karsiyaka, in the provincial centre of İzmir, Turkey, and one school from each of the three districts have been selected to constitute a random cluster sample. All the 8th-grade students of these three schools constituted student sample of the study. Student sample size comprised a total of 443 students and contained 75, 165 and 204 students from each school according to the school sizes. A student sample size of 443 was sufficient for performing the G theory analyses. Because the selection of raters who are experts in the field will increase rating reliability (Schoonen, Vergeer, & Eiting, 1997), rater sample comprised four instructors chosen among Turkish instructors who are experts in their field.

Data Collection Tool and Research Data

All the students who constituted the sample were asked to write an essay. The topic of the essay was selected from the topics provided by three Turkish teachers and two

experts of educational measurement and assessment. Furthermore, instructions on the essay topic were given as follows:

> Success is not a gift that can be obtained because of coincidence but is a product of a certain amount of hard work. It is a victory that is achieved because of planned and determined work. The key to being successful is not to work for several hours but to work in a planned manner. Those who hold this key have no alternative but to succeed. Based on this explanation, write an essay explaining the importance of planned work

According to the abovementioned instructions, students wrote their essays in their own schools during Turkish class in one period (45 min) in a writing area that did not exceed the standard writing area, which can be defined as 70 lines and approximately one and a half pages of an A4-sized paper.

Furthermore, the Writing Skill Scale (WSS) (Dogan, 2015) was used for rating students' essays. This scale, which is an ordinal scale, comprised 20 items. Each item is rated on a quaternary-scale (none=0, insufficient=1, partially sufficient=2 and sufficient = 3). Because of the application of exploratory factor analysis for determining factorial construct validity, three factors with eigenvalues of greater than one were obtained. These three factors explained 82.82% of the total variance. Because of Varimax rotation, factor loads were observed to be between .74 and .87 in 14 items of the first sub-dimension, between .84 and .89 in 3 items of the second sub-dimension, and between .87 and .97 in 3 items of the third sub-dimension. These sub-dimensions were examined by experts, and the first sub-dimension was named as wording/writing (14 items), the second sub-dimension as paragraph construction (3 items), and the third sub-dimension as title selection (3 items).

Training raters with respect to rating can increase rating reliability (Weilgle, 1994; Weigle, 1998). Moreover, a good knowledge of the rating criteria affects the reliability of the ratings (Schoonen, 2005). Therefore, training was provided to four selected Turkish lesson teachers for understanding how to rate and how to use the scoring scale. Furthermore, essays to be rated were divided into four and distributed to the raters. Raters were requested to write their ratings in a separate electronic tablet that was reserved for each rater. Essays that were obtained from the raters who finished rating the essays provided to them were given to other raters. Thus, it was ensured that every rater rated all essays and that they were completely independent of each other in rating. A data matrix containing 443 × 20 dimensions was obtained because all students' papers were rated by each rater using a 20-item ordinal scale with three sub-dimensions. Furthermore, data matrices of four teachers were combined and prepared for analysis.

Data Analysis

The 20-item WSS used for rating comprised three sub-dimensions with a different number of items. Thus, sub-dimensions will be fixed facets and items will be nested in these facets. When sub-dimensions are crossed with 's', 'x' and symbolised as nested in ':', the design becomes a univariate G theory design that can be symbolised as $p \times (i:s) \times r$ because all objects of measurement (p) are rated by all raters (r) on all items (i) in each sub-dimension (s). Brennan (2001a) refers to such designs as the 'table

of specifications' designs that comprise a sub-dimension (or tests) and items in a sub-dimension. Such a design is considered to be balanced when the number of items in each sub-dimension is equal; otherwise, it is considered to be unbalanced. This study used an unbalanced design because the number of items in sub-dimensions was different. Brennan (2001a, p. 86) states in G theory that the usage of multivariate G theory analysis instead of univariate analysis in unbalanced designs, as in this study, is a more convenient and powerful methodology. Furthermore, univariate analysis creates uncertainty and complexity in estimates and designs with unequal number of items in sub-dimensions, whereas multivariate analysis ensures separate estimation of variance and covariance components in each fixed facet sub-dimension (Brennan, 2001a, p. 276); therefore, herein, a multivariate $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ design of G theory is used. In this design, superscripted and filled circle '•' denotes that the facet is crossed with fixed multivariate data, and unfilled circle '◦' denotes that the facet is nested in multivariate data (Brennan, 2001a; Brennan, 2001b).

Variance components are estimated for sub-dimensions in G study conducted using the multivariate design $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$of the G theory. Herein, the generalizability coefficient ($E\rho^2$) was calculated for performing relative measurements, and reliability coefficient Φ was calculated for performing the absolute, sub-dimension and compound measurements. In the alternative D study, the $E\rho^2$ and Φ coefficients were calculated with an increased and decreased number of rater scenarios for sub-dimensions and compound measurements. All the G theory analyses were conducted using the mGENOVA 2.1 PC (Brennan, 2001b) version software.

Results

The findings are presented below respectively in two stages which are labelled as multivariate generalizability study and multivariate decision study.

Multivariate Generalizability Study

In generalizability (G) study using the multivariate design $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ of the G theory, three main (*p*, *i* and *r*) and four interaction effect variances (*pxi*, *pxr*, *ixr* and *pxixr,e*) were estimated. These variances, which were separately estimated for the sub-dimensions and their percentages in the total variance, are presented in Table 1.

Table 1

Variances and Percentages for the Sub-dimensions Estimated using G study

	Title selection		Paragraph Construction		Wording/Writing	
Source*	Variance	%	Variance	%	Variance	%
P	.90326	73.90	1.08388	75.51	.32054	49.88
I	.06262	5.12	.00141	.10	.03129	4.87
R	.03361	2.75	.02942	2.05	.0102	1.59
p x i	.05269	4.31	.00116	.08	.0311	4.84
p x r	.08986	7.35	.24923	17.36	.08243	12.83
i x r	.00871	.71	.00192	.14	.02315	3.60
P x i x r, e	.07161	5.86	.06831	4.76	.14391	22.39
Total	1.22236	100.00	1.43533	100.00	.64262	100.00

: P: object of measurement, i: item, r: rater, e: error

Title Selection Sub-dimension. The percentage of the object of measurement (*p*) variance, which is also referred to as the universe variance, in the total variance is expected to be greater than the remaining main and interaction variances for an optimal measurement (Brennan, 2001a; Shavelson & Webb, 1991). Thus, the object of measurement variance (*p*) having the greatest variance (79.90%) in the total variance of this sub-dimension has denoted individuals' diversity with respect to 'title selection' abilities in the essays that they have written. Item (*i*) variance constituted 5.12% of the total variance. Relatively large percentage of variance associated with the items can be interpreted as differentiation of items in the "title finding" sub-dimension. The fact that another main effect variance, rater (*r*) variance, which was the focal point of this study, constituted a relatively small fraction (2.75%) of the total variance showed that there is little discrepancy among raters' ratings in the 'title selection' sub-dimension. Thus, only a few differences presumably existed in terms of generosity or rigidity in this sub-dimension with regard to ratings for all objects of measurement by four raters. The fact that variance of the interaction effect between the object of measurement and item (*pxi*), which was estimated as 4.31% of the total variance, was relatively high denoted that relative conditions of the objects of measurement differed between various items in the 'title selection' sub-dimension. Variance of interaction effect between the object of measurement and rater (*pxr*) constituted 7.35% of the total variance. This observation denotes that certain raters rated certain objects of measurement rigidly or generously in this 'title selection' sub-dimension, i.e. the relative rankings of certain objects of measurement differed for certain raters. Variance of the interaction effect between item and rater (*ixr*) constituted .71% of the total variance. The fact that the share of this variance in total variance was close to zero denoted that the raters rated students from one item to another in a consistent manner. The final variance, i.e. residual variance, comprised trilateral interaction occurring among the object of measurement, rater and item as well as error variance (*pxrxi,e* or residual). It has been concluded that relative rankings of the objects of measurement in this sub-dimension constituted 5.86% of the total variance of trilateral interaction variance of the objects of measurement, rater and item along with remaining error sources that were not taken into consideration during the G study.

Paragraph Construction Sub-dimension. Variance estimated for the object of measurement (*p*) main effect was the greatest constituting 75.51% of the total variance, denoting the diversity of the ability of 'paragraph construction' in submitted essays. Item (*i*) variance constituted .10%, which was a considerably small fraction, of the total variance. This denoted that items in the paragraph construction sub-dimension only exhibited a minor variation. The fact that rater (*r*) variance constituted a small fraction of the total variance with 2.05% denoted that there was a minor discrepancy between the ratings of the raters in the sub-dimension. Percentages of bilateral variance of the interaction effect between the object of measurement and item (*pxi*) and between item and rater (*ixr*) were .08% and .14%, respectively, and were observed to be close to zero. Thus, relative conditions of the objects of measurement among the items of this sub-dimension differed slightly, and raters rated the objects of measurement from one item to another in a consistent manner. On the contrary, variance of interaction effect between the object of measurement and rater (*pxr*) constituted 17.36% of the total

variance as the greatest variance after object of measurement (*p*) variance, which is the universal rating variance. This denoted that certain raters rated certain objects of measurement either rigidly or generously. The residual variance, which is the trilateral interaction among the object of measurement, rater and item as well as error variance (*pxrxi,e* or residual) and the variance of the relative rankings of the object of measurement, rater and trilateral interaction of item and other error sources that were not considered in the G study, were found to constitute 4.76% of the total variance.

Wording/Writing Sub-dimension. Object of measurement (*p*) variance, which is the universal rating variance, constituted a smaller percentage, 49.88%, of the total variance when compared with other sub-dimensions. However, for the object of measurement (*p*) main effect variance, which was the greatest variance in the total variance, students' 'wording/writing' ability diversity has been put forth in essays, although in a lesser degree in comparison with other sub-dimensions. Item (i) variance, which constituted 4.87% of the total variance, had a relatively higher percentage and showed differentiation of items in this sub-dimension. The fact that rater (r) variance, whichwas the focal point of this study, constituted a relatively small portion of the total variance with 1.59% and denoted that the ratings of the raters in this sub-dimension showed little discrepancy or that there were few differences in terms of generosity or rigidity. The fact that the bilateral variance of the interaction effect between the object of measurement and item (*pxi*), which constituted 4.31% of the total variance,was relatively large denoted that the relative conditions of the objects of measurement differed in this sub-dimension. The fact that bilateral variance of interaction effect between the item and rater was 3.60% of the total variance denoted that rating stability of the raters while rating the objects of measurements between various items was lower when compared to that observed in other sub-dimensions. Bilateral variance of interaction effect between the object of measurement and rater (*pxr*) constituted 12.83% of the total variance. This denoted that certain raters rated certain objects of measurement either rigidly or generously. The residual variance, which occurred because of trilateral interaction among the object of measurement, rater and item as well as error variance (*pxrxi,e* or residual), exhibited the second greatest variance percentage, i.e. 22.39% of the total variance. This indicated that relative rankings of the objects of measurement exhibited a great variance of trilateral interaction among the object of measurement, rater and item, which was larger than the remaining error sources that were not considered in the G study.

Multivariate Decision Study

In the decision study (D) with a multivariate design $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ of the G theory, the G and Phi coefficients were calculated for four raters of the original study and for higher and lower number of raters as alternatives in each sub-dimension and in the compound measurement. Different number of raters and the G and Phi coefficient estimates for sub-dimensions and compound ratings are presented in Table 2.

Table 2

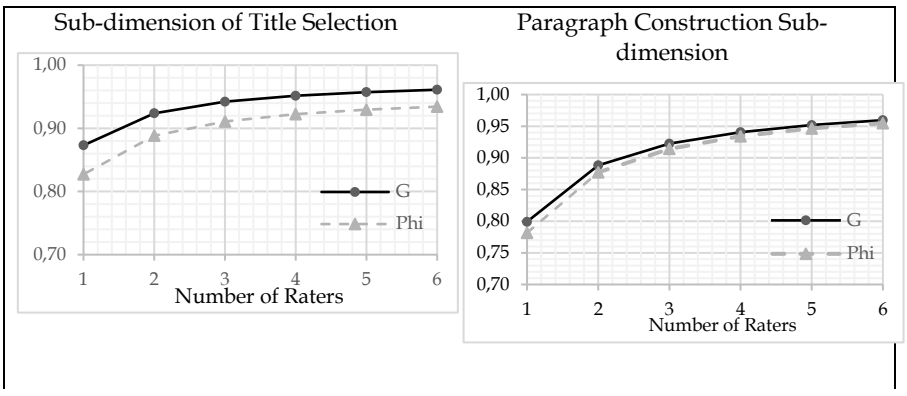
G and Phi Estimates for Different Number of Raters

Number of Raters	Sub-dimensions						Compound Measurement	
	Title Selection		Paragraph Construction		Wording/ Writing			
	G	Phi	G	Phi	G	Phi	G	Phi
6	.96	.93	.96	.96	.95	.94	.97	.96
5	.96	.93	.95	.95	.94	.93	.96	.95
4	*.95*	*.92*	*.94*	*.93*	*.93*	*.91*	*.95*	*.94*
3	.94	.91	.92	.91	.91	.89	.94	.93
2	.92	.89	.89	.88	.87	.85	.91	.90
1	.87	.83	.80	.78	.77	.75	.84	.82

Note: The italicised figures are the original number of raters.

The G coefficient ($E\rho^2$), which is calculated for the norm-referenced measurements, was obtained for the four raters as .95, .94 and .93 for 'title selection', 'paragraph construction' and 'wording/writing', respectively, and as .95 for compound measurement. The Φ coefficient, which measures the reliability of absolute (criterion-referenced) measurements, was calculated for the four raters who provided the ratings in the study as .92, .93 and .91 for the sub-dimensions of 'title selection', 'paragraph construction' and 'wording/writing', respectively, and as .94 for compound measurement.

The D study was conducted using different number of raters to determine the effect of the number of raters on the generalizability and reliability (dependability) of essay ratings, to determine the manner in which variances of the number of raters changed the G and Phi coefficients and to determine the optimal number of raters with the G theory perspective by considering manpower, time and economy without compromising psychometric quality. The effect of the number of raters obtained in the D study on the G and Phi coefficients for the sub-dimensions and compound measurement are presented in Figure 1.



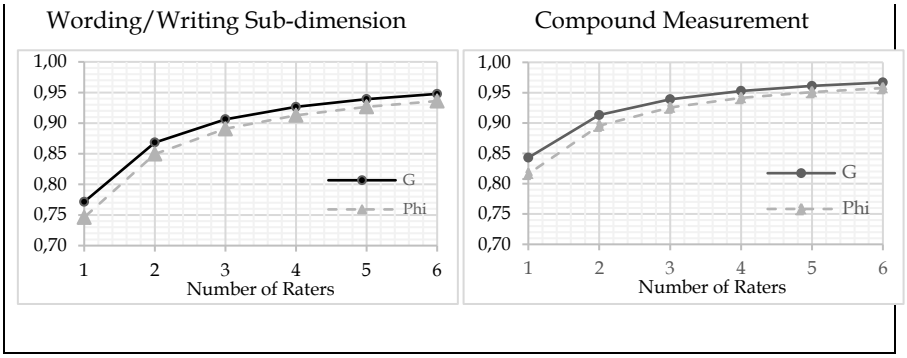


Figure 1. Sub-dimension and Compound Measurement G and Phi Coefficients for Different Number of Raters

Herein, four randomly selected Turkish course instructors were trained on how to rate students' essays. As presented in Table 2, all the G and Phi coefficients for sub-dimensions and compound measurement were greater than .90 with respect to the ratings of these four raters. As depicted in Figure 1, when the number of raters was increased from four to five, there was little gain in the G and Phi coefficients for sub-dimensions and compound measurement; when the number of raters was reduced to three, there was very little loss, and all the coefficients remain greater than .89. However, when the number of raters was reduced to two, there was some increase in the loss of the G and Phi coefficients for sub-dimensions and compound measurement. At the same time, in case two raters provided the rating, obtained G coefficients were .92, .89 and .87 and the Phi coefficients were .89, .88 and .85 for the sub-dimensions of 'title selection', 'paragraph construction' and 'wording/writing', respectively. As can be observed from Figure 1, when three raters instead of two provided the rating, the gain obtained decreased in the sub-dimensions and, particularly, in the compound measurement.

Discussion, Conclusion and Recommendations

One of the aims of this study was to determine the effect of raters on reliability. Therefore, ratings of raters, who were experts in their fields and who were trained on how to rate the essays and how to use the scale of rating, were analysed. In the G study, although wording/writing sub-dimension was smaller than title selection and paragraph construction sub-dimensions, the calculated variance of the object of measurement exhibited the highest share. The main effect variances of the raters were observed to be relatively small in the sub-dimensions, and this observation showed that ratings given for all the objects of measurement by the trained raters were consistent with each other. This result is similar to the findings of several previously conducted studies (Kim, Schatschneider, Wanzek, Gatlin & Otaiba, 2017; Schoonen R., 2005; Sudweeks, Reeve & Bradshaw, 2005) with respect to rating of writing abilities in the literature and shows that rater variance is small and that raters provide ratings

consistently with each other. Simultaneously, the fact that the percentage of the variance of interaction effect between item and rater (*ixr*) was small in all sub-dimensions can be attributed to raters being consistent in rating the items. Furthermore, people who will provide ratings should be chosen from relevant experts (Schoonen, Vergeer & Eiting, 1997) and should be trained; in these trainings (Weilgle, 1994; Weigle, 1998), they should be taught how to rate, and should also understand that provision of rating criteria affects the reliability of ratings (Schoonen, 2005). However, high percentage of variances of interaction effect between the object of measurement and rater (*pxr*) shows that certain raters were either rigid or generous in rating certain objects of measurement in all sub-dimensions. These results indicated that trained raters, who can provide consistent ratings for all objects of measurement and items, may rate a certain object of measurement more rigidly or generously and may not show the same level of consistency with respect to relative rankings of the objects of measurement. This situation (Schoonen, 2005) supports the view that even trained raters often cannot come to an agreement on rating. In this context, considering this topic while training experts will be appropriate to reduce variance between the object of measurement and rater (*pxr*) and to prevent differences between the ratings of certain raters. Moreover, with an increase in the experience of trained raters, this problem will decrease.

Another objective of this study was to establish a reference for providing future essay ratings by determining the optimal number of raters with respect to manpower, time and economy without compromising on the psychometric quality. In the analyses of the K study using the multivariate design $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ of the G theory, the G and Phi coefficients were observed to be .95 and .94, respectively, for compound measurement among which the original coefficients were obtained using four raters, and these coefficients were observed to be greater than .90 and high in all the sub-dimensions. An increase in the number of raters with alternative K studies provided little gain in the coefficients that were obtained with four trained raters and that were observed to be already high. At the same time, when the number of trained raters was one, the G and Phi coefficients of compound measurement were obtained as .84 and .82, respectively; furthermore, when the number of trained raters was two, the G and Phi coefficients of the compound measurement were obtained as .91 and .90, respectively. This result is consistent with the finding of Kim, Schatschneider, Wanzek, Gatlin and Otaiba (2017), who suggested that one rater and several tasks are required to achieve a reliability of .80 and that two raters and several tasks are required to achieve a reliability of .90.

The results of this study suggest that two raters who are trained on the subject of rating will ensure that the G and Phi coefficients are greater than .90 while rating essay writing abilities of students. In this study, a crossed design was used. However, because a significant amount of time is required for all raters to rate all the persons, particularly in large scale tests, further research should be conducted using different designs as an alternative to crossed designs, such as nested design, by allowing some raters to rate some persons.

References

Atilgan, 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. https://doi.org/10.1080/17437270801919925.

Atılgan, H., Kan, A., & Aydin, B. (2017). *Egitimde olcme ve degerlendirme [Measurement and evaluation in education].* Ankara: Anı Yayıncılık.

Baykul, Y. (2000). *Egitimde ve psikolojide olcme: Klasik Test Teorisi ve uygulamasi [Measurement in education and psychology: Classical Test Theory and application].* Ankara: OSYM.

Bereiter, C. (2003). Foreword. In M. D. Shermis, & J. C. Burstein (Ed.), *Automated essay* (pp. 7-9). NJ: LEA: Mahwah.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement*, *22*, 41-52. https://doi.org/10.1111/j.1745-3984.1985.tb01048.x.

Bouwer, R., Beguin, A., Sanders, T., & van den Berg, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1) 83-100. https://doi.org/10.1177/0265532214542994.

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

Brennan, R. L. (2001b). *Manual for mGENOVA (version2.1). Iowa Testing Programmes, Occasional Papers Number 50.* Iowa City, IA: University of Iowa.

Cohen, R. J., Swerdlik, M. E., & Philips, S. M. (1996). *Pyschological testing and assessment: An inroduct on to test and measurement* (3th Edition). California: Mayfield Publishing Company.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research.* Princeton, NJ: Educational Testing Service. GRE Board Research Report GREB No. 82-15R=ETS Research Report 84-12.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart, and Winston.

Cronbach, J. L. (1984). *Essentials of psychological testing.* New York: Happers&Row Publishers.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York, NY: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of Generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*, 67–96. https://doi.org/10.1111/1540-4781.00137.

Doğan, N. (2015). Yazılı yoklamalar [Written examinations]. In H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (pp. 145-168). Ankara: Anı Yayıncılık.

Doğan, C. D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, *7*, 631–651. http://dx.doi.org/10.12738/estp.2017.2.0321.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater. *Language Testing*, *25*, 155–185. https://doi.org/10.1177/0265532207086780.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*, 507–531. https://doi.org/10.1177/0265532209340188.

Graham, S. (2006). Writing. In P. Alexander, & P. Winne (Ed.), *Handbook of educational psychology* (pp. 457-478). NJ: Erlbaum: Mahwah.

Graham, S., Harris, K., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment. A carnegie Corporation Time to Act report.* Washington, DC: Alliance for Excellent Education.

Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalizability theory. *Perceptual and Motor Skills*, *105*, 891-905. https://doi.org/10.2466/pms.105.3.891-905.

Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Otaiba, S. A. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Read Writ*, *30,* 1287-1310.

Kondo-Brown, K. (2002). A facets analysis of rater bias in measuring Japanese second language writing. *Language Testing*, *19,* 3–31. https://doi.org/10.1191/0265532202lt218oa.

Lord, F., & Novick, M. (1968). *Statistical Theory of mental test score.* California: Addison-Wesley Publishing Company.

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

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

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

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.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Berg (Ed.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1-22). Leiden, The Netherlands: Brill.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*, 157-84 https://doi.org/10.1177/026553229701400203.

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

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 239-261. https://doi.org/10.1016/j.asw.2004.11.001.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Education and Psychological Measurement*, *59*, 492–506. https://doi.org/10.1177/00131649921970008.

Thordike, L. R. (1971). *Educational measurement (2nd. Edition).* Washington: American Council on Education.

Tugut, F. (1995). *Egitimde olcme ve degerlendirme metodları [Measurement and evaluation methods in education].* Ankara: Nüve Matbaası.

Turgut, M., & Baykul, Y. (2010). *Egitimde olcme ve degerlendirme [Measurement and evaluation in education].* Ankara: Pegem Akademi.

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

Weilgle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 197-223. https://doi.org/10.1177/026553229401100206.

Kompozisyon Puanlarının Güvenirliği: Genellenebilirlik Kuramı Çalışması

Atıf:

Atilgan, H. (2019). Reliability of essay ratings: A study on generalizability Theory. *Eurasian Journal of Educational Research*, *80*, 133-150, DOI: 10.14689/ejer.2019.80.7

Özet

Problem Durumu: Kompozisyonların puanlanmasında puanlayıcılar arasındaki bu farklılaşmaların ölçme hatalarına neden olması puanların tutarsızlığı ve güvenirliğin düşmesi ile sonuçlanır. Kompozisyon tipi sınavlarla ölçülen becerilerin ne derece doğrulukla puanlanabildiğinin belirlenmesi varyans kaynaklarından gelen ölçme hatalarının ortaya konulmasına bağlıdır. Aynı zamanda ölçme sonuçlarına karışan bu tür ölçme hatalarının azaltılması için de bu hata kaynaklarının doğru şekilde bilinmesi ve ölçme durumunun ona göre desenlenmesi gerekir.

Araştırmanın Amacı: Bu Çalışmada çok değişkenli G Kuramı kapsamında bireylerin Türkçe anadilde yazma becerilerin ölçüldüğü kompozisyon puanlarının genellenebilirliği ve güvenirliği incelenmiştir. Bu bağlamda kompozisyon puanlarının genellenebilirliği ve güvenirliği üzerine yukarıda belirtildiği gibi daha etkili olduğu bilinen puanlama konusunda eğitilmiş puanlayıcıların etkisi üzerine odaklanılmıştır. Örtük özellik olan yazma becerisinin tümü ve alt boyutları olarak başlık bulma, paragraf oluşturma, anlatım-yazma boyutlarında puanlamada puanlayıcı sayısına göre genellenebilirlik ve güvenirlik katsayılarının nasıl değiştiği ve optimal bir güvenirlik için en uygun puanlayıcı sayısının ne olabileceği ortaya konulmaya çalışılmıştır. Böylece kompozisyon yazma becerilerinin değerlendirilmesi konusunda bilgimizi genişletmek ve kompozisyonların yeterince güvenilir puanlanması için referans oluşturmak amaçlanmıştır.

Araştırmanın Yöntemi: Çalışmada kullanılan okul örneklemi; Türkiye'de İzmir il merkezinden önce üç ilçe, sonra bu üç ilçenin her birinden birer okul yansız küme örneklem olarak seçilmiştir. Örnekleme seçilen okulların 8. sınıf öğrencilerinin tamamı öğrenci örneklemini oluşturmuştur. Öğrenci örneklemi 443 öğrenciden oluşmaktadır. Puanlayıcı örneklemi ise konusunda uzman olan Türkçe dersi öğretmenleri arasından seçilen dört öğretmenden oluşturulmuştur. Öğrencilerin kompozisyonlarını puanlamak için Yazma Becerileri Ölçeği (YBÖ) kullanılmıştır. Dereceleme ölçeği olan bu ölçekte 20 madde bulunmaktadır. Her bir madde dörtlü dereceleme ölçeği şeklinde puanlanmaktadır. Dört puanlayıcının kompozisyonların tümünü birbirlerinden bağımsız puanlamaları sağlanmıştır. Araştırmada G Kuramının çok değişkenli $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ deseni kullanılmıştır. G Kuramının $p^{\bullet} x\, i^{\circ} x\, r^{\bullet}$ çok değişkenli deseniyle uygulanan G çalışmasında varyans bileşenleri alt boyutlar için kestirilmiştir. Araştırmada bağıl ölçmeler için Genellenebilirlik katsayısı ($E\rho^2$), mutlak ölçmeler için güvenirlik katsayısı (Φ) alt boyutlar ve birleşik ölçme için hesaplanmıştır. Alternatif D

çalışması ile Eρ^2 and Φ katsayıları puanlayıcı sayısının artırılması ve azaltılması senaryoları ile alt boyutlar ve birleşik ölçme için hesaplanmıştır.

Araştırmanın Bulguları: G Kuramının $p^{\bullet}x\, i^{\circ}x\, r^{\bullet}$ çok değişkenli deseni Genellenebilirlik (G) çalışması ile her bir alt boyut için üç ana (*p*, *i*, *r*) ve dört ortak etki varyansı (*pxi*, *pxr*, *ixr*, *pxixr,e*) kestirilmiştir. Başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutunda birey (*p*) sırasıyla %73.90, %75.51 ve %49.88 olarak hesaplanan varyanslar toplam varyanslar içindeki en büyük varyansa sahiptir. Bu sonuç bireylerin yazdıkları kompozisyonlarda "başlık bulma" beceri farklılıklarının ortaya konulabildiğini göstermektedir. Başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutlarının madde (*i*) varyansı toplam varyansları sırasıyla %5.12, %0.10 ve %4.87 olarak bulunmuştur. Paragraf oluşturma alt boyutu dışında nispeten büyük olan bu varyans yüzdesi; başlık bulma ve anlatım/yazma alt boyutunda maddelerin farklılaştığı biçiminde yorumlanabilir. Bu çalışmanın odak noktası olan puanlayıcı (*r*) varyansı başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutunda toplam varyansın sırasıyla %2.75, %2.05 ve %1.59 olarak hesaplanmıştır. Toplam varyansların nispeten küçük bir kısmını oluşturan puanlayıcı varyansları; puanlayıcıların alt boyutunda puanlamaları arasında tutarsızlıklarının az olduğunu göstermektedir. Başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutları için kestirilen birey ve madde (*pxi*) ortak etkisi toplam varyansların sırasıyla %4.31, %.08 ve %4.84'üdür. Başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutlarında varyansların nispeten büyük oluşu, bireylerin bu alt boyutunda maddeler arasında bağıl durumlarının farklılaştığını göstermektedir. Birey ve puanlayıcı (*pxr*) arasındaki ortak etkisi varyansı başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutlarında toplam varyansın sırasıyla %7.35, %17.36 ve %12.83'ünü oluşturmaktadır. Bu sonuç alt boyutlara belli puanlayıcıların belli bireyler için daha katı ya da daha cömert puanlama yaptıklarını göstermektedir. Madde ve puanlayıcı (*ixr*) arasındaki ortak etki varyansı başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutlarında toplam varyansın %.71, %.14 ve %3.60'ı olarak hesaplanmıştır. Başlık bulma ve paragraf oluşturma alt boyutlarında bu varyansların toplam varyansları içindeki payının sıfıra yakın olması, puanlayıcıların öğrencileri bir maddeden diğerine kararlı puanladıkları biçiminde yorumlanabilirken, anlatım/yama alt boyutunda aynı kararlığın olmadığını göstermektedir. Birey, puanlayıcı, madde arasında üç yönlü ortak etki ile hata varyansları (*pxrxi,e*) başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutlarında toplam varyansların %5.86, %4.76 ve %22.39'u olarak kestirilmiştir. Alt boyutlarda, özellikle anlatım/yazma alt boyutunda büyük olan bu varyanslar bireylerin bağıl konumlarının; birey, puanlayıcı, madde üç yönlü ortak etki varyansının G çalışmasında hesaba katılmayan diğer hata kaynaklarının büyüklüğünü göstermektedir.

G Kuramının $p^{\bullet}x\, i^{\circ}x\, r^{\bullet}$ çok değişkenli deseni Karar (D) Çalışması ile her bir alt boyut ve bütün ölçek için G ve Phi katsayıları çalışmanın orijinalinde puanlama yapan dört puanlayıcı için ve alternatif olarak daha az ve daha çok puanlayıcı sayıları için hesaplanmıştır. Bağıl ölçmeler için hesaplanan G katsayısı (Eρ^2) çalışmada puanlama yapan dört puanlayıcı için "başlık bulma", "paragraf oluşturma" ve "anlatım/yazma" alt boyutları için sırasıyla .95, .94, .93 birleşik ölçme için ise .95 olarak elde edilmiştir.

Mutlak ölçmeler için puanların güvenirliğinin bir ölçüsü olan Phi (Φ) katsayısı çalışmada puanlama yapan dört puanlayıcı için "başlık bulma", "paragraf oluşturma" ve "anlatım/yazma" alt boyutları için sırasıyla .92, .93, .91 ve birleşik ölçme için ise .94 olarak hesaplanmıştır. Puanlayıcı sayısının beş puanlayıcıya çıkarılması alt boyutlar ve birleşik ölçme için G ve Phi katsayılarında çok az kazanç sağladığı gibi, üç puanlayıcıya indirildiğinde ise kayıp çok az olmakta ve tüm katsayılar .89 ve üzerinde olmaktadır. Puanlayıcı sayısı ikiye indirildiğinden alt boyutlar ve birleşik ölçme için G ve Phi katsayılarında kayıp biraz daha artmakta ancak başlık bulma, paragraf oluşturma ve anlatım/yazma alt boyutları için sırasıyla G katsayıları .92, .89, .87; Phi katsayıları .89, .88, .85 ve birleşik ölçme için G katsayısı .91, Phi katsayısı .90 olarak elde edilmektedir.

Araştırmanın Sonuç ve Önerileri: Yapılan G çalışmalarında başlık bulma, paragraf oluşturma anlatım/yazma alt boyutlarında hesaplanan birey varyansı da en büyük paya sahiptir. Puanlayıcı ana etkisi varyansları alt boyutlarda göreli olarak küçük bulunmuştur. Bu sonuç literatürde yazma becerilerinin puanlanmasına ilişkin pek çok çalışmada puanlayıcı varyansının küçük ve puanlayıcıların birbirleri ile tutarlı puanlamalar yaptıkları bulguları ile benzerdir. Madde ve puanlayıcı (*ixr*) arasındaki ortak etki varyansı yüzdesinin tüm alt boyutlarda küçük olması puanlayıcıların maddeleri puanlamada tutalı oldukları şeklinde yorumlanabilir. Elde edilen bu sonuçlar puanlama yapacak kişilerin puanlama yapacakları konunun uzmanlarından seçilmesi, eğitilmesi ve bu eğitimlerde neyin nasıl puanlanması gerektiği, puanlama kriterlerinin verilmesi durumunda puanların güvenirliğinin yüksek olacağını göstermiştir. Ancak birey ve puanlayıcı (*pxr*) ortak etki varyansı yüzdelerinin tüm alt boyutlarda yüksek oluşu belli puanlayıcıların belli bireyleri puanlamalarında daha katı ya da cömert olduklarını göstermektedir. Bu bağlamda birey ve puanlayıcı (*pxr*) arasındaki ortak etki varyansının küçültülebilmesi ve böylece belli puanlayıcıların belli bireyleri puanlamalarında katılık ya da cömertlik bakımından farklılıkların olmaması için kompozisyon puanlayacak uzmanların eğitiminde bu konunun dikkate alınması yerinde olacaktır. Ayrıca puanlama yapacak uzman ve eğitimli puanlayıcıların puanlama deneyimlerinin artması ile bu sorunun da azalacağı düşünülebilir.

K çalışması analizlerinde, orijinali dört puanlayıcıyla yürütülen çalışmada birleşik ölçme için G katsayısının .95 ve Phi katsayısının .94 olduğu, tüm alt ölçeklerde bu katsayıların .90'ın üzerinde ve oldukça yüksek olduğu görülmüştür. Alternatif K çalışmaları ile puanlayıcı sayısının artırılması uzman ve eğitimli dört puanlayıcı ile elde edilen katsayılarda çok az kazanç sağlamıştır. Bununla birlikte uzman ve eğitimli puanlayıcı sayısının iki olması durumunda ise G katsayısı .91, Phi katsayısı .90 olarak elde edilmiştir. Bu sonuç .90 üzerinde bir güvenirliğe ulaşmak için iki puanlayıcının yeterli olduğunu göstermiştir.

Anahtar Kelimeler: Genellenebilirlik Kuramı, genellenebilirlik, güvenirlik, kompozisyon puanlama, kompozisyon puanlama güvenirliği, puanlayıcı güvenirliği, yazma puanlaması.