Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 14(2) (2022) 1629–1643

IJCI
International Journal of Curriculum and Instruction

Comparison of the results of the generalizability theory with the inter-rater agreement coefficients

Mehmet Taha Eser [a] *, Gökhan Aksu [b]

[a,b] *Aydın Adnan Menderes University, Faculty of Education, Aydın, 09010, Turkey*

Abstract

The agreement between raters is examined within the scope of the concept of "inter-rater reliability". Although there are clear definitions of the concepts of agreement between raters and reliability between raters, there is no clear information about the conditions under which agreement and reliability level methods are appropriate to use. In this study, the comparison of eight different agreement coefficients used for the same purpose and the similarity of the results obtained with the G coefficient calculated within the framework of generalizability theory were examined. Within the scope of the study, it was determined that there were differences between the agreement coefficients of the evaluations made by the seven raters for 49 students over six open-ended items. As a result of the study, it was determined that the agreement coefficients differed significantly according to the method used and the level of agreement could be interpreted as low-medium-high according to the method used. In addition, as a result of the generalizability analysis, it was determined that the largest proportion of the variance components resulted from the difference between the raters and equal to 40% of the total variance between the raters. For this reason, it is recommended that researchers first examine the variance components originating from the person, item, and raters while determining the inter-rater reliability, and finally, report a few of the appropriate coefficients in case the inter-rater variance is low.

Keywords: Inter-rater agreement; inter-rater reliability; generalizability theory

1. Introduction

1.1. Introduce the problem

More than one person is trusted to collect data in studies conducted within the scope of science. The problem of consistency, or agreement between individuals collecting data, arises immediately because of variability among raters. For this reason, well-designed research studies should include some procedures to measure agreement between

* Corresponding author: Mehmet Taha Eser. ORCID ID.: https://orcid.org/0000-0001-7031-1953
E-mail address: m.taha.eser@adu.edu.tr

different raters. Study designs include subjecting data collectors to specific training and measuring the extent to which these individuals record the same scores for the same individual/situation/object. It is rarely seen that the agreement between the raters is perfect, that is, there is a perfect agreement between the raters. The agreement between raters is examined within the scope of the concept of "inter-rater reliability".

Rater reliability is the degree of consistency between the scores given in general terms. If this consistency is sought between the scores given by a rater, it is expressed as intra-rater reliability, and if it is considered based on the compatibility between the scores of more than one rater, it is expressed as inter-rater reliability (Johnson, Penney & Gordon, 2000). Inter-rater reliability, expressed in various ways such as evaluator reliability, observer reliability, and rater reliability, is the degree of agreement or consistency between two or more raters (Cohen et al, 1996). Inter-rater reliability focuses on whether the student's score changes from rater to rater, and it is taken into account that the rater may have subjective judgments in the scores. In a study with raters, reliability turns into reliability between raters, in other words, the amount of relationship or agreement between two or more coders (Neuendorf, 2002). It is important that the value obtained as a result of the inter-rater reliability calculation reaches an acceptable level, since it does not have the practical advantages of establishing the basic nature of the scoring plan and scoring multiple raters.

If there is a clear definition of the concept of inter-rater reliability, it is thought that the usefulness of the statistics used in the calculation of inter-rater reliability can be discussed. However, there is no information in the literature that there is a clear definition of the concept of inter-rater reliability. Some coefficients, such as intraclass correlation coefficients, are based on variance decomposition, which is in harmony with the environment related to generalizability theory (Vangeneugden et al., 2005). Coefficients such as the percentage of agreement are derived by considering the concept of literal agreement. Not all coefficients that define inter-rater reliability based on different conceptualizations can measure the same behavior. Krippendorff (2016), who recently had a discussion with Feng (2015) on rater reliability, claims that Feng has seriously misunderstood that reliability criteria should ensure that we are assured of reliability. A more accurate terminology is needed to distinguish between the different theories underlying inter-rater reliability coefficients and to define competing conceptualizations of inter-rater reliability. If the theories and models underlying the concept of inter-rater reliability can be expressed much more clearly, we can begin to investigate why some of the inter-rater reliability coefficients produce higher or lower values than other coefficients. In the context of expressing the theories and models underlying the concept of inter-rater reliability much more clearly, the work of Zhao et al. (2013) makes a very important contribution to the literature. Zhao et al. (2013) mentioned the limitations of chance-corrected coefficients such as kappa (κ). Within the scope of the same study, it is noted that the biggest difference between the reliability

coefficients between the raters is due to the "rater behavior" (the coefficients included in the percentage of agreement assume that the raters never code randomly, while the agreement coefficients based on chance assume that the raters maximize random coding).

Inter-rater reliability can be calculated in any case when there are two or more independent raters measuring the same object. Inter-rater reliability is a measure of the level of agreement between independent coding performed by two or more raters (Hallgren, 2012). The reliability value obtained reflects the extent to which the raters agree on the scoring of a certain behavior (Burry-Stock et al., 1996). With inter-rater reliability, the consistency of the coding is determined and information is obtained about how much a coder's choices deviate from the ideal or "correct" codes. Results from scoring by different raters or at different time points (for example, before and after an intervention) are of great importance in many disciplines where abilities, behaviors, and symptoms are so frequently evaluated and compared. Experts working in the fields of behavioral psychology and education emphasize that it is necessary to clearly determine the points where many different statistics contribute to the evaluation of the fit and reliability of the categories used within the scope of rater reliability (Mitchell, 1979; Stemler, 2004). The precise definitions and distinctions made regarding the concepts contribute to preventing the misleading interpretation of the data. There are many different statistics that can measure inter-rater reliability. Among these very different statistics that measure inter-rater reliability, Cohen's kappa coefficient is the best known and most widely used by researchers.

Most researchers associate inter-rater reliability with Cohen's (1960) kappa. Although Cohen's kappa seems to be a "symbol coefficient" by researchers regarding inter-rater reliability, it has many limitations. Kappa cannot be applied to non-categorical data. Since kappa is affected by sample size, it is very difficult to compare kappa values obtained from different substances or different studies. Kappa is designed to include chance agreement in the calculation, but its assumptions about rater independence and other factors are not sufficient. For these reasons, the kappa can be calculated as lower than the required value in some cases.

When the literature is examined, it is striking that there are many coefficients for inter-rater reliability. For nominal data only, Popping (1988) mentions that there are more than 38 coefficients. Zhao et al. (2013) discussed 22 of these coefficients, determined that several of the 22 related coefficients were mathematically equivalent, and concluded that there were 11 unique coefficients. In addition, within the scope of “irr”, an R software package, it includes 17 different coefficients for various data types that predict inter-rater reliability (Gamer et al. 2012). As a result of the different versions of some coefficients between raters, the number of coefficients increases even more. To give an example, one-way intra-class correlation coefficient model or two-way intra-class correlation coefficient model is used to estimate the consistency or consistency of a single

scoring performed within the scope of interclass correlation coefficient (ICC) or the mean between raters. Due to the large number of coefficients, it seems very difficult for one coefficient to be more prominent than another at the point of choosing a certain coefficient to estimate the reliability between raters. Although the strengths of the coefficients regarding rater reliability are emphasized in the literature, it is still not clear to what extent the estimated inter-rater reliability depends on the coefficient (Hallgren 2012; Gwet 2014)

Pearson's product-moment correlation coefficient is one of the most used coefficients to determine the fit between raters. In order to use this correlation technique, the scores must be continuously variable and at least on an equally spaced scale. However, the values obtained show the variability of the scores given by the raters and are insufficient to explain the reliability. Therefore, Burry-Stock et al. (1996) stated that the correlation coefficient explains the co-variability of the scores, but is far from showing the agreement between the raters. Different methods and techniques are needed due to the limitations of the correlation technique and other similar methods such as parallel forms, test-retest method, split-half method. These are Cohen's Kappa, weighted Kappa, Kendall's W, Krippendorff's Alpha, Scott's phi, Holsti method, Lin's concordance correlation coefficient, Cochran's Q test, Logistic regression, Loglinear analysis etc. . can be listed as In addition, variance analysis is also used in determining the consistency between raters, in cases where the data is a continuous variable. Technically, intra-class correlation coefficient is calculated based on analysis of variance (Bıkmaz, 2011). Within the scope of the research, Kendall's W, Pearson's correlation coefficient, Iota coefficient, Finn coefficient, ICC, Brennan-Prediger coefficient, Gwet's AC_1 coefficient and Krippendorff's alpha coefficient were used to calculate the agreement coefficient between raters. Below is a very brief introductory information about the relevant coefficients.

Pearson's product-moment correlation coefficient is one of the most used coefficients to determine the fit between raters. In order to use this correlation technique, the scores must be continuously variable and at least on an interval scale. However, the values obtained show the variability of the scores given by the raters and are insufficient to explain the reliability. Therefore, Burry-Stock et al. (1996) stated that the correlation coefficient explains the co-variability of the scores, but is far from showing the agreement between the raters. Due to the limitations of the correlation technique and other similar methods such as parallel forms, test-retest method, split-half method, different methods and techniques (Cohen's Kappa, weighted Kappa, Kendall's W, Krippendorff's Alpha, Scott 's phi, Holsti method, Lin's concordance correlation coefficient, Cochran's Q test, Logistic regression, Loglinear analysis etc.) are needed. In addition, variance analysis is also used in determining the consistency between raters, in cases where the data is a continuous variable. Technically, intra-class correlation coefficient is calculated based on analysis of variance (Bıkmaz, 2011). Within the scope of the research, Kendall's W, Pearson's correlation coefficient, Iota coefficient, Finn coefficient, ICC, Brennan-Prediger

coefficient, Gwet's AC_1 coefficient and Krippendorff's alpha coefficient were used to calculate the agreement coefficient between raters. Below is a very brief introductory information about the relevant coefficients.

Kendall's W coefficient of agreement is a non-parametric statistic used to evaluate agreement between three or more raters and takes a value in the range of 0-1. Within the scope of the Iota coefficient (Janson & Olssons, 2001), an extension of Cohen's (1960) kappa, chance-corrected agreement is calculated for a multivariate test scored by two or more raters. Finn agreement coefficient is a coefficient used to determine inter-rater reliability in cases where the opinions of the raters are stated as quantitative data. The intraclass correlation coefficient can be defined as the ratio of the variance to each other in the results obtained from the subjects in general. Brennan and Prediger (1981) recommend using the agreement coefficient when there are two raters and an arbitrary number of q categories. Holley and Guilford (1964) were the first to use the Brennan-Prediger agreement coefficient to calculate reliability and defined this coefficient as the G-index. Krippendorff's Alpha (α) statistic, on the other hand, is a agreement coefficient that can be applied to a wide variety of data types and any number of values for each variable, and can also be used in cases where there is missing data (Krippendorff, 1995). Gwet's AC_1 coefficient is a agreement coefficient developed to eliminate the limitations and shortcomings of Cohen's kappa. The value obtained as a result of calculating the AC_1 coefficient of Gwet is for crossed designs (Gwet, 2008).

When it comes to inter-rater reliability, there are opinions in the literature that some popular and frequently used statistics are insufficient in calculating inter-rater reliability. For example, the Cronbach alpha was designed only to measure internal consistency and to standardize the variance of assessments made by different raters (Hughes & Garrett, 1990). Chi-square produces high values for both agreement and disagreement that deviate from random agreement ("expected values in the chi-square formula"). Before making a final decision on the method of determining inter-rater reliability, the characteristics of the raters, the assumptions of the chosen method, the scale type of each variable for which the agreement will be calculated, the number of raters, and the characteristics of the data should be taken into account.

Classical test theory methods cannot be applied directly if raters rank individuals' traits of interest. In such cases, variance analysis techniques are used to identify the sources of variation. Generalizability (G) theory is a statistical theory that enables the evaluation of reliability in measuring behavior, designing, researching and conceptualizing reliable observations, and is based on analysis of variance (ANOVA) (Shavelson & Webb, 1991). Item, time, rater and similar error sources are considered as sources of variance within the scope of G theory. Within the scope of the research, G theory was used to calculate the reliability between raters.

The aim of the study is to determine the level of agreement between the results obtained by seven independent raters by using the analytical rubric of student performances in the Statistics and Probability course written exam prepared for the same purpose and administered to the same individuals. For this purpose, the greement values obtained with Kendall's W, Pearson's correlation coefficient, Iota coefficient, Finn coefficient, ICC, Brennan-Prediger coefficient, Gwet's AC_1 coefficient and Krippendorff's alpha coefficient were examined and compared. At the same time, in this study, in order to determine the reason for the difference in the agreement coefficients, based on the generalizability theory, the level of difference between the individual, the item and the raters, which are the components of the variance, was examined.

2. Method

2.1. Research design

This research is based on the application of different agreement coefficients, determining the similarities and differences of these techniques, examining their limitations, and determining which of the techniques provides more information. In this respect, it is a descriptive research since it is aimed at determining the situation.

2.2. Study group

The study group of the research consists of 49 volunteer students who took the 2019-2020 Fall Term Statistics and Probability Course and 7 volunteer experts who scored the answers given by these students to the Statistics and Probability Course Written Exam questions consisting of 6 open-ended questions. The students in the study group were from Aydın Adnan Menderes University, Faculty of Education.

2.3. Data collection tools

In many studies on education and psychology, independent raters are required to score in order to measure some behavioral characteristics. For example, raters can be used to score open-ended tasks on a standardized test, rate the performance of expert athletes in a sporting event, or experimentally test the applicability of a new rubric. Scoring processes in these examples are processes in which objective scoring for the behavior, which is the subject of measurement, cannot be realized. Considering this situation, within the scope of the research, an analytical rubric was used to score the answers given by the students.

2.4. Data analysis

In the analysis of the data, the agreement coefficients between the raters and the G theory were used for the analytical rubric. Within the scope of the research, the R

programming language was used in the calculation of the agreement coefficients between the raters, and the EduG program was used in the calculation of the G coefficient.

3. Results

The agreement coefficients calculated for the 8 different methods obtained for the first item scored between 1 and 5 within the scope of the study are shown in Table 1.

Table 1. Comparison of agreement coefficients regarding the first item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.69	-	0.000	-	-
2. Pearson	0.68	-	0.000	-	-
3. Iota	0.64	-	-	-	-
4. Finn	0.91	-	0.000	-	-
5. ICC	0.93	-	0.000	0.90	0.95
6. Brennan	0.41	0.04	0.000	0.33	0.49
7. Gwet AC_1	0.42	0.04	0.000	0.34	0.50
8. Krippendorf α	0.31	0.05	0.000	0.21	0.40

When Table 1 is examined, it is seen that the highest agreement coefficient for the 1st item in the measurement tool is ICC, while the lowest agreement coefficient is the Krippendorff alpha coefficient. In addition, among the coefficients obtained, Kendal's W coefficient, Pearson's correlation coefficient and Iota coefficient give close results; Brennan-Prediger agreement coefficient and Gwet's AC_1 coefficient give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive.

The agreement coefficients calculated for the 8 different methods obtained for the second item scored between 1-10 within the scope of the study are shown in Table 2.

Table 2. Comparison of agreement coefficients regarding the second item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.95	-	0.000	-	-
2. Pearson	0.94	-	0.000	-	-
3. Iota	0.88	-	-	-	-
4. Finn	0.77	-	0.000	-	-
5. ICC	0.98	-	0.000	0.97	0.99
6. Brennan	0.55	0.06	0.000	0.43	0.66
7. Gwet AC_1	0.56	0.06	0.000	0.44	0.67
8. Krippendorf α	0.46	0.04	0.000	0.38	0.54

When Table 2 is examined, it is seen that the highest agreement coefficient for the 2nd item in the measurement tool is ICC, while the lowest agreement coefficient is the

Krippendorff alpha coefficient. In addition, among the coefficients obtained, Kendall's W, Pearson's correlation coefficient and ICC give close results; Brennan-Prediger coefficient and Gwet's AC_1 coefficient give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive.

The agreement coefficients calculated for the 8 different methods obtained for the third item scored between 1 and 5 within the scope of the study are shown in Table 3.

Table 3. Comparison of agreement coefficients regarding the third item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.71	-	0.000	-	-
2. Pearson	0.72	-	0.000	-	-
3. Iota	0.62	-	-	-	-
4. Finn	0.96	-	0.000	-	-
5. ICC	0.92	-	0.000	0.89	0.95
6. Brennan	0.78	0.05	0.000	0.69	0.88
7. Gwet AC_1	0.81	0.04	0.000	0.72	0.89
8. Krippendorf α	0.47	0.06	0.000	0.34	0.60

When Table 3 is examined, it is seen that the highest agreement coefficient for the 1st item in the measurement tool is Finn, while the lowest agreement coefficient is the Krippendorff alpha coefficient. In addition, while Kendall's W coefficient and Pearson's correlation coefficient are close to each other; Brennan-Prediger coefficient and Gwet's AC_1 coefficient give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive.

The agreement coefficients calculated for the 8 different methods obtained for the fourth item scored between 1 and 5 within the scope of the study are shown in Table 4.

Table 4. Comparison of agreement coefficients regarding the fourth item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.87	-	0.000	-	-
2. Pearson	0.84	-	0.000	-	-
3. Iota	0.73	-	-	-	-
4. Finn	0.89	-	0.000	-	-
5. ICC	0.95	-	0.000	0.92	0.97
6. Brennan	0.37	0.05	0.000	0.28	0.46
7. Gwet AC_1	0.37	0.05	0.000	0.28	0.47
8. Krippendorf α	0.32	0.05	0.000	0.22	0.42

When Table 4 is examined, it is seen that the highest agreement coefficient for the 4th item in the measurement tool is ICC, while the lowest agreement coefficient is the Krippendorff alpha coefficient. In addition, among the coefficients obtained, Kendall's W

coefficient, Pearson's correlation and Finn coefficients give close results; Brennan-Prediger, Gwet's AC_1 and Krippendorff alpha coefficients give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive.

The agreement coefficients calculated for the 8 different methods obtained for the fifth item scored between 1 and 8 within the scope of the study are shown in Table 5.

Table 5. Comparison of agreement coefficients regarding the fifth item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.93	-	0.000	-	-
2. Pearson	0.87	-	0.000	-	-
3. Iota	0.84	-	-	-	-
4. Finn	0.89	-	0.000	-	-
5. ICC	0.97	-	0.000	0.96	0.98
6. Brennan	0.58	0.05	0.000	0.49	0.68
7. Gwet AC_1	0.59	0.05	0.000	0.50	0.68
8. Krippendorf α	0.53	0.05	0.000	0.43	0.62

When Table 5 is examined, it is seen that the highest agreement coefficient for the 5th item in the measurement tool is ICC, while the lowest agreement coefficient is the Krippendorff alpha coefficient. In addition, among the coefficients obtained, Kendall's W coefficient, Pearson's correlation coefficient and Finn coefficient give close results; Brennan-Prediger coefficient, Gwet's AC_1 coefficient and Krippendorff alpha coefficient give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive.

The agreement coefficients calculated for the 8 different methods obtained for the sixth item, scored between 1 and 10, are shown in Table 6.

Table 6. Comparison of agreement coefficients regarding the sixth item

Method	Estimate	Std. Error	p	Ci Lower Bound	Ci Upper Bound
1. Kendall W	0.88	-	0.000	-	-
2. Pearson	0.97	-	0.000	-	-
3. Iota	0.82	-	-	-	-
4. Finn	0.81	-	0.000	-	-
5. ICC	0.97	-	0.000	0.96	0.98
6. Brennan	0.35	0.03	0.000	0.28	0.41
7. Gwet AC_1	0.35	0.03	0.000	0.28	0.42
8. Krippendorf α	0.30	0.03	0.000	0.23	0.37

When Table 6 is examined, it is seen that the highest agreement coefficient for the 6th item in the measurement tool is ICC, while the lowest agreement coefficient is the

Krippendorff alpha coefficient. In addition, while the correlation coefficients of ICC and Pearson, which are among the coefficients obtained, are close to each other; Brennan-Prediger, Gwet's AC_1 and Krippendorff alpha coefficients give close results. However, it is seen that the Iota and Finn coefficients give close results. In addition, it is seen that all the agreement coefficients are statistically significant and the agreement coefficients are positive. The comparison of the agreement coefficients calculated within the scope of the study is shown in Table 7.

Table 7. Comparison of results

Method	Value Range	Item1 (5p)	Item3 (5p)	Item4 (5p)	Item5 (8p)	Item6 (10p)	Item2 (10p)
1. Kendall W	[0, 1]	0.69	0.71	0.87	0.93	0.88	0.95
2. Pearson	[-1, 1]	0.68	0.72	0.84	0.87	0.97	0.94
3. Iota	[0, 1]	0.64	0.62	0.73	0.84	0.82	0.88
4. Finn	[0, 1]	0.91	0.96	0.89	0.89	0.81	0.77
5. ICC	[0, 1]	0.93	0.92	0.95	0.97	0.97	0.98
6. Brennan	[0, 1]	0.41	0.78	0.37	0.58	0.35	0.55
7. Gwet AC_1	[0, 1]	0.42	0.81	0.37	0.59	0.35	0.56
8. Krippendorf α	[0, 1]	0.31	0.47	0.32	0.53	0.30	0.46

When Table 7 is examined, it is seen that the Krippendorf alpha coefficient, which is one of the agreement coefficients in the same value range, has the lowest value in all six items in the test, although it is used for the same purpose. In addition, it was determined that Brennan-Prediger coefficient and Gwet's AC_1 coefficient were close to each other in almost all items and were the second coefficients that gave the lowest value. When the values of Kendal's W coefficient are examined, it is seen that the agreement coefficient increases when the points given by the experts widen. It is seen that Kendal's W coefficient gives close results with Pearson and ICC, especially in items evaluated over 8-10 points. It was determined that the mean of the Pearson agreement coefficients was related to the score range of the relevant item, and the Pearson agreement coefficient increased when the score range expanded. Similarly, it was determined that the Iota agreement coefficient increased when the score range expanded. On the contrary, it was determined that the Finn coefficient decreased when the score range was expanded. It was observed that the ICC coefficient was not affected much by the score range, and it was the coefficient with the highest value for all items. According to these results, it has been determined that other agreement coefficients, except ICC, are affected by the score range of the items. It was also determined that while the ICC coefficient for each item gave the highest value, Krippendorff's alpha tended to have the lowest value.

As mentioned before, G theory was used to calculate inter-rater reliability within the scope of the research. Table 8 contains the results of the G study conducted through the EduG program.

Table 8. Analysis results of variance components

Source	SS	df	MS	Random	Mixed	Corrected	%	SE
S	1959.98	48	40.83	0.71	0.71	0.71	9.40	0.19
R	5715.81	6	952.63	3.04	3.04	3.04	39.70	1.62
I	230.88	5	46.17	-0.00	-0.00	-0.00	0.00	0.08
SR	3144.79	288	10.91	1.59	1.59	1.59	20.90	0.15
SI	259.92	240	1.08	-0.03	-0.03	-0.03	0.00	0.01
RI	1470.86	30	49.02	0.97	0.97	0.97	12.70	0.25
SRI	1909.65	1440	1.32	1.32	1.32	1.32	17.30	0.04
Total	14691.94	2057					%100	

*S : Students, R : Raters, I : Items, SR : Students x Raters, SI : Students x Items, RI : Raters x Items

When Table 8 is examined, only 9.40% of the total variance is due to the difference between students, while the amount of variance resulting from the difference between raters is 39.70%. It is seen that the variance arising from the items is 0.0%. This result can be interpreted as there is no difference in the difficulty levels of the items, in other words, the items are on at the same difficulty level. It is expected that the highest value among the variance components given in Table 8 will originate from the students. However, the biggest variance here is due to the difference between the raters. Similarly, the calculation of the student-rater interaction variance component as 20.90% indicates that there are differences between the raters in terms of stinginess and generosity. Although each of the raters gave points to the students according to the analytical rubric given to them, this result shows that the raters gave different results among the students. The rater-item interaction variance component was calculated as 12.70% and this value shows that there may be differences between the scores given by the raters according to the items. In addition, it was determined that the Relative G coefficient was 0.73 and the Absolute G coefficient (phi) was 0.50.

4. Discussion and Conclusions

In this study, the comparison of eight different agreement coefficients used for the same purpose and the similarity of the results obtained with the G coefficient calculated within the framework of generalizability theory were examined. Within the scope of the study, it was determined that there were differences between the agreement coefficients of the evaluations made by the seven raters for 49 students over six open-ended items. In particular, while Krippendorff's alpha coefficient tended to have the lowest value in all items, it was determined that ICC and later Kendall's W and Pearson coefficients were relatively higher than other coefficients. In addition, it was concluded that the agreement coefficients were also affected by the score ranges for the items.

The questions that need to be answered while making a decision on which or which of the inter-rater reliability and inter-rater agreement coefficients will be used can be examined under three headings. The first of these is about determining the scales of the

measurement (nominal, ordinal, interval, ratio). The second question should focus on the number of raters. Finally, the question "Should the raters agree completely or what should be the lowest acceptable rate of agreement as long as the differences are systematic?" must be answered.

There is no coefficient called "best" or "ideal" in the calculation of reliability and agreement coefficients between raters. In this sense, each coefficient has some advantages and disadvantages besides having assumptions. An example of this is that the raters' probability of giving the same score due to the chance factor affects the value of the agreement coefficient.

When the studies in the literature are examined in general, it is striking that the researchers preferred simple statistical methods to complex statistical methods in terms of calculating the reliability and agreement coefficients between the raters. Basic methods may yield the results that may be needed, but more advanced computational methods may provide more complementary statistics on these results. Considering the results obtained with the R program within the scope of this study, only the level of agreement was determined within the scope of the Iota coefficient; It can also be determined whether the level of agreement regarding the Kendall, Pearson and Finn coefficient is statistically significant. On the other hand, confidence intervals for the level of agreement can also be calculated in Brennan-Prediger coefficient, Gwet's AC1 coefficient and Krippendorff's alpha coefficients.

Within the scope of this study, ICC, which has the highest agreement coefficients, is one of the most used methods to determine the agreement between raters. However, interpreting the results only according to the ICC in inter-rater agreement studies has various drawbacks. Since ICC is a method that measures relative agreement, it cannot distinguish between systematic error and random error (Atkinson & Nevill, 1998; Weir, 2005). For this reason, it is of great importance to use absolute reliability methods and standard error of measurement (SEM), which are not affected by the variability in measurement values, together with ICC in order to prevent misinterpretation of ICC values (Weir, 2005). As a matter of fact, as a result of determining the G coefficient as 0.73 and the variance due to the rater component as approximately 40% in this study, it would be wrong to make the comment that "Inter-rater reliability is high" considering only the ICC. Similarly, it would be wrong to comment that "the agreement between raters is low" by considering only Krippendorff's alpha coefficient. For this reason, researchers are recommended to first examine the individual, item, rater and the variance components resulting from their interaction while interpreting the coefficients related to the level of agreement between raters, and then report a few of the agreement coefficients together if the variance components between raters are low. As a matter of fact, based on the findings of this study, among the agreement coefficients that vary in the range of 0-1, if only ICC is used, the agreement between the raters is high; When only

Krippendorff's alpha coefficient is used, it should be noted that the level of agreement is low. For this reason, it is recommended that researchers report other agreement coefficients before making assertive comments based on Krippendorf's alpha coefficient, which tends to give relatively low values.

Although within the scope of this study, seven different raters were given an analytical scoring key and scored, as a result of the generalizability analysis, it was determined that the largest variance component originated from the raters. For this reason, the level of agreement between the raters can be increased and more reliable measurement results can be obtained by giving training to the raters or by creating clear and understandable instructions before the analyzes regarding the determination of the agreement coefficient. In addition, within the scope of the pilot study, the evaluations can be compared by having the raters apply, and thus, the results can be made more consistent and high level of agreement coefficients can be obtained.

References

Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Atkinson G., & Nevill A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.*, *26*(4): 217-38.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, *27*(1), 3-23.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(1) , 63-78 . DOI: 10.21031/epod.294847

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687-699.

Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assess. Writ. 9*, 105–121. **https://doi.org/10.1016/j.asw.2004.07.001**

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater agreement indexes for performance assessment. *Educational and Psychological Measurement*, *56*(2), 251–262. **https://doi.org/10.1177/0013164496056002006**

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423-429.

Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography. *Journal of Personality and Social Psychology*, *70*(5), 945–960. **https://doi.org/10.1037/0022-3514.70.5.945**

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of the two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543- 549.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 13–22. **https://doi.org/10.1027/1614-2241/a000086**

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23-34.

Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement*, *24*(4), 749–753. **https://doi.org/10.1177/001316446402400402**

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research*, *27*, 185-195.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*, 277-289.

Johnson, R. L., Penney, J., & Gordon, B. (2000). The relation between score resolution methods and ınterrater reliability: a

An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(2), 121-138.

Kraemer, H. C. (1979). Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*, *44*(4), 461– 472

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, *25*, 47-76.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *12*(4), 139–144. https://doi.org/10.1027/1614-2241/a000119

Maclure, M., & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, *126*(2), 161-169.

Matthias Gamer, Jim Lemon and Ian Fellows Puspendra Singh <puspendra.pusp22@gmail.com> (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*(2), 376.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Oleckno, W. (2008). *Epidemiology: Concepts and methods*. Waveland Press, Inc.

Popping, R. (1988). *On agreement indices for nominal data*. In W. E. Saris & I. N. Gallhofer (Eds.), Sociometric research (pp. 90-105). London, UK: Palgrave Macmillan. doi:10.1007/978-1-349-19051-5_6

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

Stemler, S. E. (2004). A Comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *57*(7), 959-972.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231-40.

Xie, Q. (2013). Agree or disagree? A demonstration of an alternative statistic to Cohens kappa for measuring the extent and reliability of agreement between observer. In Proceedings of the Federal Committee on Statistical Methodology Research Conference, The Council of Professional Associations on Federal Statistics, Washington, DC, USA.

Zhao, X. (2011). When to use Cohen's K, If ever?, International Communication Association 2011 Conference, Boston, Massachusetts, U.S.A.

Zhao X., Liu J. S., & Deng K. (2013) Assumptions behind inter-coder reliability indices. Annals of the International *Communication Association*, *36*(1), 419-480.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*(3), 374-378.