Measuring Essay Assessment:

Intra-rater and Inter-rater Reliability

Ulaş KAYAPINAR[*]

Suggested Citation:

Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136 http://dx.doi.org/10.14689/ejer.2014.57.2

Abstract

Problem Statement: There have been many attempts to research the effective assessment of writing ability, and many proposals for how this might be done. In this sense, rater reliability plays a crucial role for making vital decisions about testees in different turning points of both educational and professional life. Intra-rater and inter-rater reliability of essay assessments made by using different assessing tools should also be discussed with the assessment processes.

Purpose of Study: The purpose of the study is to reveal possible variation or consistency in grading essay writing ability of EFL writers by the same/different raters using general impression marking (GIM), essay criteria checklist (ECC), and essay assessment scale (ESAS), and discuss rater reliability.

Methods: Quantitative and qualitative data were used to present the discussion and implications for the reliability of ratings and the consistency of the measurement results. The assessing tools were applied to 44 EFL university students and 10 graders assessed the essay writing ability of the students by using GIM, ECC, and ESAS in different occasions.

Findings and Results: The findings and results of the analyses indicated that using general impression marking is evidently not reliable for assessing essays. The coefficients obtained from checklist and scale assessments, considering the correlation coefficients, estimated variance components, and generalizability coefficients present valuable information, clearly show that there is always variation among the results.

[*] Dr., EPP, American University of the Middle East, Kuwait. ukayapinar@gmail.com

Conclusions and Recommendations: When the total scores and the rater consensus results in this study are examined, it can be clearly seen that the scores are almost always not identical and they are different from each other. For this reason, opposed to the idea that is commonly agreed upon, checklists or even scales may not be effectively as reliable as expected and they may not improve inter-reliability or intra-reliability of ratings unless the raters are very well-trained and they have strong agreement or common inferences on performance indicators and descriptors since they should not have ambiguous interpretations on the criteria set. The results might be more accurate and reliable if the accepted interpretation of a meaningful correlation coefficient for this kind of measurements can be considered as .90 minimum for giving evidence of reliable ratings. This might mean that the proximity of the scores which are assigned to same or independent essays will be higher and more similar. However, the scale use could still be emphasized as more reliable. Still, an elaborate and careful examination with more raters is seen needed.

Keywords: Essay, assessment, intra-rater, inter-rater, reliability.

Assessing writing ability and the reliability of ratings have been a challenging concern for decades and there is always variation in the elements of writing preferred by raters and there are extraneous factors causing variation (Blok, 1985; Chase, 1968; Chase, 1983; Darus, 2006; East, 2009; Engelhard, 1994; Gyagenda & Engelhard, 1998a; Gyagenda & Engelhard, 1998b; Hughes, Keeling & Tuck, 1980; Hughes, Keeling & Tuck, 1983; Hughes & Keeling, 1984; Kan, 2005; Klein & Hart, 1968; Klein & Taub, 2005; Marshall & Powers, 1969; Murphy & Balzer, 1989; Schaefer, 2008; Slomp, 2012; Sulsky & Balzer, 1988; Wexley & Youtz, 1985; Woehr & Huffcutt, 1994). Fisher, Brooks, and Lewis (2002) state fitness for purpose requirement is the core of all testing work, and direct writing assessments are subjective and thereby more prone to reliability issues. For this reason, many raters use scoring scales or rubrics because they believe that any assessment without a scale is based on subjective judgments and general impression. Some researchers also state that not only general impression marking but also holistic assessment with a set of criteria can be highly subjective (Hamp-Lyons, 1991; Vaughan, 1991) and scores can vary in a significant way. Huot (1990) states that the levels of reliability achieved with holistic assessment are generally lower than that achieved with analytic assessment (Johnson, Penny, & Gordon, 2001). In this respect, general impression marking and holistic assessment can be called as subjective but analytic assessment can be called objective-like or systematically subjective because, all in all, each indicator of criteria is scored subjectively (Kayapinar, 2010). Even if it seems more reliable than the others, there is still a set of criteria which is implicit or explicit for different types of assessment. Moreover, a comparison of reliability measures by using different assessment tools is seen necessary in order to provide evidence going beyond any claim and reaching the proof of assessing essays consistently because the rating methods –holistic or analytic- used by the raters can change their application of rating criteria (Huang, 2012).

In this article, general impression marking refers to handling with an essay as a whole with a subjective judgment (Hamp-Lyons, 1992). For this reason, no tool was addressed for this type of assessment in the study. Holistic assessment refers to scoring the overall product as a whole, with judging the predetermined component parts separately (Mertler, 2001; Nitko, 2001). For this type of assessment, a checklist entitled Essay Criteria Checklist (App.1) was employed. A rating scale entitled Essay Assessment Scale (App.2) was used for analytic assessment which refers to scoring the levels of the product with individual predetermined criteria and obtaining a total score by the sum of the individual scores (Moskal, 2000; Nitko, 2001; Weir, 1990).

Considering the measures of rater reliability and the carry-over effect, the basic research question guided in the study is in the following:

Is there any variation in intra-rater reliability and inter-reliability of the writing scores assigned to EFL essays by using general impression marking, holistic scoring, and analytic scoring?

Method

Sample

Three study groups were randomly chosen and employed as follows: *Judges.* Judges (*n*=103) include faculty of ELT departments from different (20) universities. They evaluated the appropriateness and validity of the checklist items (App. 1) and the criteria and performance indicators of the scale (App. 2). *Raters.* Raters (*n*= 10) who assessed the essays are ELT experts (MAs and PhDs) and experienced teachers of writing skill (at least 2 years). *EFL students.* The students (*n*= 44) who responded the essay test produced the essays in testing conditions for Advanced Reading and Writing class.

Research Instruments

The writing samples. Forty-four scripts of one essay sample written in testing conditions in order to achieve the objective :

"By means of the awareness of essay types, essay writers will analyze, synthesize and evaluate information and therefore, in their compositions, react to prompts. Essay writers will also be able to analyze and produce different types of essays (e.g. comparison and contrast, classification, process analysis, cause-and-effect analysis, and argumentative) that are unified, coherent, and organized." The essay prompt, which was produced by the teachers of the particular class, is the same for all students as: ***Please write an essay about the topic "University students should be free to choose their own courses."***

Essay Criteria Checklist (ECC). The checklist was developed in order to measure each construct of essay writing. First of all, a criteria list was written through a review of relevant literature (Raimes, 1983; Norton, 1990; Celce-Murcia, 2001; Johnson, Penny, & Gordon, 2001; Jacobs *et al*. 1981 in Weigle, 2002; Weigle, 2002;

Bowen and Cali, 2004; Hawkey & Barker, 2004; Darus, 2006; IELTS, 2007; Dempsey, PytlikZillig, & Bruning, 2009; Knoch, 2009). Next, 103 faculty from ELT departments from different (20) universities examined the appropriateness of the checklist considering the expressions used and the consistency between the objectives and constructs of essay writing skill and the checklist items. The ratio of agreement (P) (Erkuş, 2003) was found significantly high (P=96.1; P= the number of judges agreed on each criterion/total number of judges). Later, two experts of measurement and evaluation examined the checklist considering the content and technical features.

Essay Assessment Scale (ESAS). The scale was developed in order to describe and measure each construct of essay writing skill with performance levels. First, 103 faculty of ELT departments from different (20) universities examined the scale considering the expressions used and the consistency between the objectives and constructs of essay writing skill and the performance indicators included. The ratio of agreement (P) of the scale is also .96.1. Next, two experts of measurement and evaluation examined the scale considering technical features. Finally, a Likert type scale covering five performance levels (0-1-2-3-4) was developed by using expert judgments. Five performance levels were chosen because of easiness and usefulness for the observable behavior although there is no limit for performance levels (Kan, 2007).

The measurement results: The total scores of 2640 ((10 raters × 44 essay scripts) × 6 independent sessions) essay scripts, which were randomly selected, were used to measure the reliability of ratings, using GIM, ECC, and ESAS.

S*tandardized open-ended interviews.* The raters were asked the following standardized open-ended interview questions about the assessment process:

1.“What do you think of the assessments you made by using GIM?”

2.“What do you think of the assessments you made by using ECC?”

3.“What do you think of the assessments you made by using ESAS?”

A pretest of the interview questions was carried out by two independent raters and two experts of measurement and evaluation in order to identify the validity and the effectiveness of the questions.

Procedure

The procedure of the study includes two phases: *The production of the material to be scored*. The essays were produced in testing conditions of an advanced reading and writing class. Each essay was given a different code assigned randomly for each rating after the names had been deleted.

Assessment Design. There are ten raters and six different rating processes in the study. Before the raters started each rating session, they had been given a short educational session and instructions for a proper completion of each session. Each rater scored each essay at a time -44 essays in one batch and 264 essays in total. Each rating session was held after a 10-week break in order to remove the carry-over effect of the previous assessment. In order to balance the objectivity, the order and the numbering of the essays were changed before each session and they were assigned random codes.

Data Analyses

In order to determine the intra-rater reliability of the ratings, the correlation coefficients between the two gradings of the same raters for the same essays were computed by using Pearson Product Moments Correlation Analysis. The correlation coefficients were also examined by using Fischer's z Transformation to test the significance of the variation in correlation coefficients. This procedure led the way to put the correlation coefficients in order. ANOVA was employed in order to present evidence for the inter-rater reliability of ratings. The differences in the scores across the task and the raters by using GIM and ESAS were also interpreted through a generalizability study. A series of person × rater × task were performed to examine the variation of scores due to potential effects of person, rater, and task after the variance components had been estimated. Using standardized open-ended interviews revealed the reflections and views of the raters on their own rating process. The qualitative data here were analyzed line by line and memos were written (Glesne, 1999; Strauss & Corbin, 1998). Categories were reviewed and recurring themes, core consistencies and meanings were identified by using pattern codes. Those explanatory pattern codes were later identified as smaller sets and themes with content analysis (Miles & Hubermas, 1994; Patton, 2002). The process includes: Underlying key terms in the responses, restating key phrases, coding key terms, pattern coding, constructing themes, and corporating themes into an explanatory framework

Results

Intra-rater reliability.

Table 1 shows the intra-rater consensus between GIM assessments.

Table 1

Intra-rater Consensus between GIM Assessments

Difference	*R1*		*R2*		*R3*		*R4*		*R5*		*R6*		*R7*		*R8*		*R9*		*R10*	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%
0	7	16	2	5	6	14	6	14	1	2	1	2	1	2	9	21	7	15	3	7
±1-5	9	21	17	38	8	18	18	41	13	30	30	68	10	23	7	15	1	3	18	41
±6-10	8	18	7	15	9	21	8	18	13	30	12	27	6	14	7	15	7	15	14	32
±11-15	9	21	6	14	8	18	4	9	6	14	0	0	6	14	9	21	2	5	2	5
±15- more	11	25	12	27	13	30	8	18	11	25	1	2	21	47	12	27	12	27	7	15
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

R=Rater

Table 1 shows that Rater 6 scored 31 essays out of 44 with a ±0-5-point difference on 0-100 point scale. This is the highest value among the others referring that 70% of the essays have similar results in two assessments made by using GIM. The assessments of Rater 9 have the lowest percentage of consensus which is 18% with a ±0-5-point difference. The frequency is 7 for zero difference, and 1 for ±1-5-point difference. Other raters' consensus between two assessments by using GIM has a frequency range between 11 and 21 points. Table 2 also indicates that the percentages of the scores which are the same in two assessments have a range between 2 and 21. This means that the frequencies range between 1 and 9 out of 44 essays. Rater 5, 6, and 7 have only one score which is the same for both assessments. However, Rater 8 scored 9 essays the same. For a better understanding of the rater reliability of general impression marking, it is necessary to examine the correlation coefficients between the two assessments made by using GIM. The correlation coefficients computed, by using Pearson Product Moments Correlation, are presented below in Table 2:

Table 2

Correlations across GIM Assessments

Rater	*r*
1	.042
2	.510**
3	.477**
4	.279
5	.450**
6	.835**
7	.584**
8	.412**
9	.790**
10	.880**

** Correlation is significant at the 0.01 level

The correlation coefficients, seen in Table 2, range between .042 and .880. Among the ten coefficients, two of them, which belong to the raters 1 and 4, are not significant. The other correlation coefficients seem significant. This may mean that those raters assigned similar scores to the essays in both assessments. However, only 3 of them are above .70 which refers to a considerably high and meaningful correlation (Kline, 1986) and relatively a high consistency. In fact, even the coefficient of .70 seems insufficient for a high level of consistency when the intra-rater consensus is examined and the results in Table 1 and 2 are compared carefully. For example, Rater 10 scored only 3 essays (7%) with no difference and 18 essays (41 %) out of 44 with a ±1-5-point difference in spite of the highest correlation coefficient obtained (.880) among GIM assessments.

Table 3

Intra-rater Consensus Between ECC Assessments

Difference	*R1*		*R2*		*R3*		*R4*		*R5*		*R6*		*R7*		*R8*		*R9*		*R10*	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%
0	1	2	1	2	1	2	3	7	6	14	1	2	5	11	0	0	1	2	2	5
±1-5	9	21	35	80	14	32	14	32	36	82	33	75	9	21	40	91	30	69	21	48
±6-10	12	27	8	8	7	16	14	32	2	5	10	23	9	21	3	7	12	27	17	39
±11-15	6	14	0	0	10	23	11	25	0	0	0	0	11	25	0	0	1	2	4	9
±15-more	14	32	0	0	12	27	2	5	0	0	0	0	36	82	1	2	0	0	0	0
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

R=Rater

Table 3 shows that Rater 5 scored 42 essays out of 44 with a ±0-5-point difference on 0-100 point scale although there are 6 essays scored with a zero difference. This is the highest value among the others referring that 96% of the essays have closer results to each other in two ECC assessments. The assessments of Rater 1 have the lowest percentage of consensus which is 23% with a ±0-5-point difference. The frequency is also 1 for zero difference, and 9 for ±1-5-point difference. Other raters' consensus between two assessments by using ECC has a frequency range between 15 and 40 points. Table 4 also indicates that the percentages of the scores which are the same in two assessments have a range between 2 and 14. This means that the frequencies range between 1 and 6 out of 44 essays. Rater 8 has no score which is the same for two assessments and the raters 1, 2, 3, 6, and 9 have only one score which is the same for two assessments. However, Rater 5 scored 6 essays the same. For a better understanding, it is necessary to examine the correlation coefficients between the two assessments made by using ECC. The correlation coefficients computed, by using Pearson Product Moments Correlation, are presented below in Table 4:

Table 4
Correlations across ECC Assessments

Rater	*r*
1	.072
2	.953**
3	.517**
4	.457
5	.955**
6	.898**
7	.730**
8	.932**
9	.928**
10	.804**

** Correlation is significant at the 0.01 level

In Table 4, the correlation coefficients range between .072 and .932, this is relatively higher than the correlation coefficients across GIM assessments. Among the ten coefficients, only one of them, which belong to the scores assigned by the rater 1, is not significant. The other correlation coefficients seem significant. This may mean that those raters gave similar scores to the essays in both assessments. However, 7 of them are above .70 which refers to a high and meaningful correlation coefficient and relatively a high consistency (Kline, 1986). Table 5 below shows the intra-rater consensus between ESAS assessments:

Table 5

Intra-rater Consensus between ESAS Assessments

Difference	*R1*		*R2*		*R3*		*R4*		*R5*		*R6*		*R7*		*R8*		*R9*		*R10*	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%
0	9	21	5	11	3	7	3	7	6	14	2	5	2	5	6	14	3	7	3	7
±1-5	18	41	21	48	18	41	3	7	28	64	18	41	24	55	24	55	23	53	24	55
±6-10	8	18	4	9	11	25	6	14	10	23	7	16	8	18	12	28	12	28	10	23
±11-15	4	9	14	32	7	16	7	16	0	0	6	14	2	5	2	5	4	9	6	14
±15-more	5	11	0	0	5	11	25	57	0	0	11	25	8	18	0	0	2	5	1	2
TOTAL	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100	44	100

R=Rater

Table 5 shows that Rater 1 scored 27 essays out of 44 with a ±0-5-point difference on 0-100 point scale. This means 62% of the essays have similar results in two assessments made by using ESAS. In the assessments of Rater 2, the number of the essays scored with ±0-5-point difference is 26, and the percentage is 59%. Rater 3 scored 21 essays with ±0-5-point difference, which means 48%. Rater 4 is the one who has the smallest amount of consistency. The rater scored only 6 essays with ±0-5-

point difference, which refers to 14%. In the assessments of Rater 5, the number of the essays scored with ±0-5 points difference is 34, which is quite high (78%) when compared to others. The results of Rater 6 show that 20 essays were scored with ±0-5-point difference on 0-100 point scale. Rater 7 scored only 2 essays the same but there are 26 essays scored with a ±0-5-point difference. Assessments of Rater 8 indicate 30 essays have ±0-5-point difference which refers to 69%. In the assessments made by Rater 9, the number of essays with ±0-5-point difference is 26. Finally, Rater 10 scored 27 essays with ±0-5-point difference with a percentage of 62. For a better understanding of the rater reliability of the scale, it is necessary to examine the correlation coefficients between the two assessments made by using ESAS. The correlation coefficients computed, by using Pearson Product- Moment Correlation, between the first and the second assessments and they are presented below in Table 6:

Table 6

Correlations across ESAS Assessments

Rater	*r*
1	.757**
2	.641**
3	.585**
4	.021
5	.825**
6	.680**
7	.545**
8	.916**
9	.811**
10	.884**

** Correlation is significant at the 0.01 level

The results indicate that the correlation coefficients between the scores raters assigned to the essays seem to be high and significant at the 0.01 level (no less than .545) except the one which was done by Rater 4 (.021). These results refer that 9 raters scored the essays in a significantly reliable way. Moreover, 7 of the correlation coefficients are around .70. This is a high level of positive correlation which is seen meaningful and which might mean that there is a high consistency between the assessments (Kline, 1986). When the results are compared to the others, Rater 4 is the one who has the smallest amount of intra-rater consistency, correspondingly, the one whose results have the lowest and the only insignificant correlation coefficient. The highest correlation coefficient belongs to Rater 8 (.916) whose scores correspond to each other. This refers to similar results for two assessments made in different time distances. Moreover, Rater 8 is the one who scored 42 essays out of 44 with ±10 points difference on 0-100 point scale (intra-rater consensus=95%). This is the best result among the raters' assessments; however, the differences among the correlation coefficients, even the ones within a 10-point difference in total scores, of the same essays scored in different times indicate there is always a source of variation in assessments made by ESAS.

Table 7

The Comparisons among Correlation Coefficients across Different Assessments

		Raters									
		1	2	3	4	5	6	7	8	9	10
The Difference between Correlation Coefficients	$r_{12} - r_{34}$	0.056	p<.05 2.433	0.099	0.016	p<.05 2.992	0.481	0.487	p<.05 2.311	1.071	0.498
	$r_{12} - r_{56}$	1.772	0.369	0.282	0.867	1.657	0.702	0.107	0.106	0.109	0.034
	$r_{34} - r_{56}$	1.648	1.572	0.176	0.849	1.282	1.137	0.141	0.205	0.961	0.531

In the table showing Fischer's z transformation, r_{12} refers to the correlation coefficient between the first two ratings; r_{34} refers to the correlation coefficient between the following two ratings; and r_{56} refers to the correlation coefficient between the final ratings. The differences at the significant level ($p<0.05$) are presented in the table. The results indicate that few raters (2, 5, and 8) made consistent and decisive assessments in different time distances. As seen in the table, no other consistent and decisive assessments were made by the raters using the same tools in different time distances. This may mean raters assign different scores to the same essays in different time distances.

Inter-rater reliability

An analysis of variance was conducted to find out the inter-rater consensus statistically. The results are given in the table below:

Table 8

Inter-rater Reliability of Assessments

Rating	Sum of Squares	df	Mean Square	F	Sig.
1	17554.036	9	1950.448	11.052	.000
2	21461.411	9	2384.601	8.913	.000
3	22407.909	9	2489.768	13.465	.000
4	20462.684	9	2273.632	10.164	.000
5	17570.475	9	1952.275	15.781	.000
6	31722.773	9	3524.753	31.983	.000

p<.0.001

The table shows the output of the ANOVA analysis and whether there is a statistically significant difference between group means. The results apparently indicate that the paired comparisons of the means of the scores raters assigned to the essays significantly differ from each other. It is clearly seen that the significance level is 0.000, which is below 0.001 (p <0.001). Therefore, there is a clear statistically significant difference in the mean scores assigned by different raters. This might mean that there are remarkable differences among scores assigned by the raters to the same essay products and the inter-rater reliability of the assessments is considerably low.

A series of a random one-facet (student × rater) model and a random two-facet model (student × task × rater) generalizability study for each rating (GIM and ESAS) were performed. It could not be realized for ECC ratings because of data loss. In addition, the generalizability study could be held for 9 raters as one of the raters was not able to provide the data for it as well. Estimated variance components for the ratings are given in Table 9 below:

Table 9

Estimated Variance Components (EVC) for GIM and ESAS ratings

Source	***n***	***GIM***		***ESAS***	
		EVC	***Total Variance %***	***EVC***	***Total Variance %***
Student	44	0.258	0.87	0.547	1.39
Task	2	2.241	7.52	3.023	7.69
Rater	9	20.215	67.86	25.951	66.05
Student × Rater		1.429	4.80	2.255	5.74
Student × Task		0.207	0.69	0.317	0.81
Task × Rater		1.989	6.68	2.556	6.51
Student × Task × Rater		3.452	11.59	4.642	11.81
Generalizability Coefficient		0.26		0.57	

In Table 9, the universal score variance increased from 0.87% to 1.39%. This reflects slight differences between those two. The s × t interactions effect seems reduced from 67.86% to 66.05% and the s × r interaction seems increased from 4.80% to 5.74%. Slightly higher variance was obtained for differences in examinees' performance across tasks when the raters assigned scores by using GIM. Besides, the s × t interaction reduced from 6.68% to 6.51% when the raters assigned scores by using ESAS. However, a pretty higher generalizability coefficient was obtained when the scores were assigned using the scale. Moreover, the s × t × r interaction increased from 11.59% to 11.81 %. This might mean that inter-rater reliability is more effective and advantageous for revealing the differences in quality of students' responses when the scale is used to assign scores to the task.

Standardized Open-ended Questioning

Standardized open-ended questioning was employed for the instrumentation of the qualitative data in order to reveal the views of the raters on assessment processes and the types of assessments. It includes the same question –the same stimuli- in the same way determined in advance (Patton, 2002). The transcripts were analyzed line by line and memos were written (Strauss & Corbin, 1998; Glesne, 1999). Categories or labels were reviewed and recurring themes, core consistencies and meanings were identified by using pattern codes (Miles & Huberman, 1994; Patton, 2002). The themes were found as : a) criteria use, b) spelling, and c) weightings

What is immediately apparent from open-ended transcripts is that the criteria use is very important and useful in essay assessment because the raters mention that they were more precise and the results were more consistent in assessing the essays by using the criteria given. One of the raters states that GIM assessments was like gambling because they needed to assign a total score to each essay without any written or pre-specified criteria. They also state that the criteria use changed the tendency of scoring subjectively in a positive manner. In this respect, raters seem to have the common idea those assessments by using a checklist or a scale is always more objective and reliable. Some teachers state that there should be a criterion for spelling. Even if the testees are advanced level writers, they might make spelling mistakes and the raters cannot score spelling because it is not one of the criteria in the scale. The spelling criterion had not been found appropriate by the judges because the task is at an advanced level. Although the raters seemed to have an agreement that GIM assessments were not reliable and consistent, they also criticized ESAS weightings. They state the criteria should not be equal for each sub-criterion. For example, one of the raters says it would be better if each weighting was different for each sub-criterion. In this way, it would be more useful and consistent. It would be particularly useful to state, considering the transcripts, that criteria use is a reliable and agreed measure for assessing essays. However, the criteria should be chosen precisely and correctly considering the needs of the students and the weightings of the criteria should be independent from each other. In fact, the weightings are different for each criterion but the particular teacher seems to think equal weightings are used for each criterion.

Discussion and Conclusions

The study gives evidence that all methods, techniques, or tools could include subjectivity and it seems reasonable to notice that mental processes and internal responses of raters function in different ways in using same assessment criteria for the same essays in different times. The statistical evidence indicates that GIM assessments are never consistent and reliable. The statistical analyses clearly show that ECC assessments are more reliable and consistent than GIM ones. The correlation coefficients are higher and they are supported by the raters themselves, as seen in qualitative data. The results also show that ESAS assessments are also consistent and reliable when compared to GIM. However, there is a slight difference between the correlation coefficients across ECC assessments and ESAS assessments. Yet, the coefficients across ESAS assessments are slightly higher and more meaningful than the ones across ECC assessments. This slight difference can also be observed by examining the intra-rater consensus between the assessments. It seems different weightings for each sub-criterion may result in more consistent assessments as raters declared because the results of the difference of correlation coefficients which were obtained by using Fischer’s z transformation also support the idea that the intra-rater scores are similar but not the same. Paired comparisons with ANOVA tell us the inter-rater scores are never meaningfully similar. This means different scores are assigned for the same essays in different time distances. It is obvious if a lower score is assigned to the same essay in two different sessions around the cut-off score, this means success and failure depend on a source of variation. At this point, the raters and the time elapsed between assessments may seem as the source of variation. The G coefficients also indicate that assigning scores is more precise and effective, when the scale is used, as it increases inter-rater reliability. Considering several limitations, further research into the effectiveness and usefulness of the scale would be valuable as it is difficult to infer what processes are experienced by the raters while they are scoring essays. The more pieces of information available, the more reliable will be the conclusions drawn from the data (Cherry & Meyer, 1993). However, when the total scores and the rater consensus results are examined, it can be clearly seen that the scores are different from each other even if the correlation coefficients are high and significant. It might be more accurate if Kline's (1986) cut-off coefficient (.70) for a meaningful correlation could be increased to .90 at least for giving evidence of more reliable ratings. This might mean the scores assigned are more similar and closer to each other. A deliberate training and agreement of raters before any process of rating for each student group also seems strongly needed on the criteria and performance indicators. In order to obtain verbal descriptions as concrete information, to recognize this process, and to establish the decision-making processes of raters, think-aloud protocols with follow-up interviews can also be employed.

References

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22*(1), 41-52.

Bowen, K. and Cali, K. (2004). *Teaching the features of effective writing.* Retrieved November 21, 2004, from http://www.learnnc.org/index.nsf/printView/1216418CB65B73CE85256D7300445C5A?OpenDocument.

Breland, H. (1983). The direct assessment of writing skill: A measurement review (Technical Report No.83-6). Princeton, NJ: College Entrance Examination Board.

Celce-Murcia, M. (2001). *Teaching English as a second or foreign language.* Massachusetts: Heinle and Heinle.

Chase, C. I. (1983). Essay test scores and reading difficulty. *Journal of Educational Measurement, 20*(3), 293-297.

Chase, C. I. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement, 2*(4), 315-318.

Cherry, R. and Meyer, P. (1993). Reliability issues in holistic assessment. In M. Williamson and B. Huot (Ed.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton.

Darus, S. (2006). Identifying dimensions and attributes of writing proficiency: development of a framework of a computer-based essay marking system for Malaysian ESL learners. *Internet Journal of e-Learning and Teaching, 3*(1), 1-25.

Dempsey, M. S., PytlikZillig, L. M., and Bruning, R. G. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14*, 38–61.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*, 88-115.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Erkuş A. (2003). *Psikometri üzerine yazilar: ölçme ve psikometrinin tarihsel kökenleri, güvenirlik, geçerlik, madde analizi, tutumlar; bileşenleri ve ölçülmesi* [Writings on Pscychometrics: historical basis for measurement and pscyhometrics, reliability, validity, item analysis, attitudes; components and measurement]. 1. baskı, Ankara. Türk Psikologlar Derneği Yayınları.

Fisher, R., Brooks, G., and Lewis, M. (2002). *Raising standards in literacy.* New York: Routledge.

Glesne, C. (1999). *Becoming qualitative researchers: An introduction.* New York: Longman.

Gyagenda, I. S. and Engelhard, G. (1998a). Rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scoring. *Annual Meeting of the American Educational Research Association.* San Diego.

Gyagenda, I. S. and Engelhard, G. (1998b). Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability. *Annual Meeting of the American Educational Research Association*. San Diego.

Hamp-Lyons, L. (1991). The writer's knowledge and our knowledge of the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (p. 15-36). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1992). Holistic writing assessment for LEP students. *Second National Research Symposium on Limited English Proficient Student Studies: Focus on Evaluation and Measurement.* Washington.

Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*, 122-159.

Herrington, A. and Moran, C. (2001). What happens when machines read our students'writing?. *College English, 63*, 480-499.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing 17*, 123-139.

Hughes, D. and Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*(3), 277-281.

Hughes, D., Keeling, B., and Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement, 20*(1), 65-70.

Hughes, D., Keeling, B., and Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17*(2), 131-135.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

IELTS (2007). *IELTS handbook*. Retrieved January 19, 2008 from http://www.ielts.org/_lib/pdf/IELTS_ Handbook_2007 .pdf#.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., and Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Johnson, R., Penny, J., and Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18*(2), 229-249.

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenirliğine etkisi [The effect of checklist and answer key use in writing assessment on rater reliability]. *Eğitim Araştırmaları Dergisi, 5*(20), 166-177.

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni program anlayışı içerisinde kullanılabilecek bir değerlendirme yaklaşımı: Rubrik puanlama yönergeleri [An evaluation approach to be used for the new curriculum considering the contributions to performance evaluation process: Rubrics]. *Kuram ve Uygulamada Eğitim Bilimleri, 7*(1), 129-152.

Kayapinar, U. (2010). A study on assessment tools and evaluation of essay writing skill in foreign language education. Unpublished PhD Dissertation, Mersin University. Yenisehir Campus: Turkey.

Klein, S. and Hart, F. M. (1968). Chance and systematic factors affecting essay grades. *Journal of Educational Measurement, 5*(3), 197-206.

Klein, J. and Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing, 10*, 134-148.

Kline, P. (1986). *A handbook of test construction: introduction to psychometric design*. London: Methuen.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275-304.

Marshall, J. C. and Powers, J. M. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement, 6*(2), 97-101.

Mertler, C. A. (2001). *Designing scoring rubrics for your classroom*. Practical Assessment, Research and Evaluation, 7(25). Retrieved October 11, 2007 from http://PAREonline.net/getvn.asp ?v=7andn=25.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis*. California: Sage Publications.

Moskal, B. M. (2000). *Scoring rubrics: what, when, and how?*. Practical Assessment, Research, And Evaluation, 7(3). Retrieved October 11, 2007 from http://pareonline.net/getvn.asp?v=7&n=3.

Murphy, K. R. and Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619-624.

Nitko, A. J. (2001). *Educational assessment of students (3rd ed.)*. Upper Saddle River, NJ: Merrill.

Norton, L. S. (1990). Essay-writing: What really counts. *Higher Education, 20*(4), 411-442.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. California: Sage Publications.

Raimes, A. (1983). *Techniques in teaching writing*. Oxford: Oxford University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing 17*, 81-91.

Strauss, A. and Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. California: Sage Publications.

Sulsky, L. M. and Balzer W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*,497-506.

Vaughan C. (1991). Holistic assessment: What goes on in the rater's mind? L. Hamp-Lyons (Ed.), *In Assessing Second Language Writing in Academic Contexts* (p. 111-126). Norwood, NJ: Ablex.

Weir, C. J. (1990). *Communicative language testing*. Wiltshire: Prentice Hall.

Wexley, K.N. and Youtz, M.A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. *Journal of Occupational Psychology, 58*, 265-275.

Woehr, D. J. and Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Appendix 1:

ESSAY CRITERIA CHECKLIST (ECC)

-Make a checkmark if the essay includes the following attributes-

CRITERIA		CHECKMARK
ORGANIZATION	A. INTRODUCTION	
	A.1.1. Introductory Sentences	
	A.1.2. Thesis Statement	
	A.2. BODY PARAGRAPHS	
	A.2.1. Topic Sentence	
	A.2.2. Supporting Sentences	
	A.3. CONCLUSION	
LANGUAGE USE	B.1. Word Order	
	B.2. Pattern Variety	
	B.3. Verb Form	
	B.4. Tenses	
	B.5. Articles	
	B.6. Pronouns	
	B.7. Prepositions	
VOCABULARY	C.1. Word Choice	
	C.2. Word Variety	
	C.3. Parts of speech	
MECHANICS	D.1. Punctuation	
	D.2. Capitalization	
	D.3. Paragraphing	
	D.4. Indentation	
IDEAS/ CONTENT	E.1. Title	
	E.2. Development	
	E.3. Unity	
	E.4. Transitional Signals	

Appendix 2:

ESSAY ASSESSMENT SCALE (ESAS)

CRITERIA		ATTRIBUTES	4	3	2	1	0
ORGANIZATION	A.1. INTRODUCTION						
	A.1.1. Introductory Sentences	Effective introductory sentences					
	A.1.2. Thesis Statement	Appropriate thesis statement (thesis and central idea)					
	A.2. BODY PARAGRAPHS						
	A.2.1. Topic Sentence	Appropriate topic sentence (possibly implied) supporting the thesis and the central idea					
	A.2.2. Supporting Sentences	Appropriate sentences supporting the topic (possibly major and minor)					
	A.3. CONCLUSION	Appropriate conclusion related to thesis					
LANGUAGE USE	B.1. Word Order	Correct word order					
	B.2. Pattern Variety	Using different patterns					
	B.3. Verb Form	Using verb forms correctly					
	B.4. Tenses	Using tenses appropriately					
	B.5. Articles	Using articles correctly					
	B.6. Pronouns	Using pronouns correctly					
	B.7. Prepositions	Using prepositions correctly (verb + preposition, adjective + preposition)					
VOCABULARY	C.1. Word Choice	Selecting the appropriate words					
	C.2. Word Variety	Having a rich vocabulary					

	C.3. Parts of speech	Using the correct parts of speech					
MECHANICS	D.1. Punctuation	Using punctuation marks correctly					
	D.2. Capitalization	Using cases (lower/upper) correctly					
	D.3. Paragraphing	Correct paragraph formatting					
	D.4. Indentation	Using margins correctly and consistently					
IDEAS/ CONTENT	E.1. Title	Appropriate title					
	E.2. Development	Appropriate development					
	E.3. Unity	Unity					
	E.4. Transitional Signals	Using appropriate transitional signals					
		TOTAL SCORE					

Kompozisyon Puanlamanın Ölçülmesi:

Aynı ve Farklı Puanlayıcı Güvenirliği

Atıf:

Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136 http://dx.doi.org/10.14689/ejer.2014.57.2

Özet

Problem Durumu: Yazma becerisinin etkili bir biçimde puanlanmasının araştırılmasına ilişkin bir hayli çaba gösterilmekte ve birçok öneri sunulmaktadır. Bu bağlamda, puanlayıcı güvenirliği, bireylerin gerek eğitim gerekse mesleki yaşamlarının farklı dönüm noktalarında hayati kararlar vermede çok önemli rol oynamaktadır. Aynı ve farklı puanlayıcıların farklı puanlama araçları kullanarak yaptıkları puanlamaların da güvenirlikleri puanlama süreçleri ile birlikte tartışılmalıdır.

Araştırmanın Amacı: Araştırmanın amacı İngilizce öğrenicilerinin yazma becerilerinin aynı/farklı puanlayıcılar tarafından genel izlenim (GIM), kontrol listesi (ECC) ve kompozisyon puanlama ölçeği (ESAS) kullanılarak değerlendirilmesindeki olası farklılık ve tutarlılıkları ortaya çıkarmak ve puanlayıcı güvenirliklerini tartışmaktır.

Yöntem: Ölçme sonuçlarının tutarlılığı ve puanlamaların güvenirliğine ilişkin yorum ve tartışmaların yapılabilmesi için nicel ve nitel veriler kullanılmıştır. Puanlama araçları 44 üniversite öğrencisi üzerinde uygulanmış ve 10 puanlayıcı genel izlenim, kontrol listesi ve ölçek kullanarak bu öğrencilerin yazma becerilerini puanlamışlardır.

Bulgular: Bulgular ve analiz sonuçları genel izlenimle puanlamanın beklendiği üzere kesinlikle güvenilir olmadığını göstermiştir. Elde edilen korelasyon katsayıları, varyans kestirimleri ve genellenebilirlik katsayilarindan elde edilen bilgiler goz onune alindiginda, puanların aynı olmadığı ve sonuçlar arasında daima bir çeşitlilik ve varyasyon olduğu görülmektedir.

Sonuç ve Öneriler: Toplam puanlar ve puanlayıcıların vermiş oldukları puanlar arasındaki tutarlılıklar incelendiğinde sonuçların, korelasyon katsayıları yüksek ve anlamlı olsa dahi, çoğu zaman aynı olmadığı ve birbirlerinden farklı oldukları görülmüştür. Bu yüzden, yaygın kanının aksine, kontrol listeleri ve ölçekler, puanlayıcıların söz konusu araçlara yönelik iyi bir eğitim almamaları ve ölçütler, ölçüt tanımları ve performans göstergeleri üzerinde bir uzlaşma sağlamadıkları takdirde beklendiği gibi etkili bir şekilde güvenilir olamayabilmektedirler.Bu tür ölçmelerde anlamlı kabul edilecek korelasyon katsayısınınn en az .90 düzeyinde olması durumunda güvenilir puanlamaya kanıt oluşturacak olan sonuçlar daha hatasız olabilir. Bu durum aynı ve farklı yazılı yoklamalara verilen puanların birbirlerine olan yakınlık düzeylerini artıracak ve daha benzer sonuçların ortaya çıkması anlamına gelebilecektir. Herşeye rağmen, hali hazırdaki durum ve sonuçlar gözönüne alındığında ölçek kullanımının diğer puanlama araçlarına göre daha güvenilir olduğu vurgulanabilir. Yine de çalışmanın daha fazla puanlayıcı ile tekrarlanmasının alana katkı sağlayacağı düşünülmektedir.

Anahtar Sözcükler: Kompozisyon, puanlama, puanlayıcılararası, puanlayıcı, güvenirlik