European Journal of Educational Research

Volume 8, Issue 1, 169 - 180.

ISSN: 2165-8714

http://www.eu-jer.com/

Reliability of the Analytic Rubric and Checklist for the Assessment of Story Writing Skills: G and Decision Study in Generalizability Theory

N. Bilge Uzun *
Mersin University, TURKEY

Devrim Alici
Mersin University, TURKEY

Mehtap Aktas
Mersin University, TURKEY

Received: October 30, 2018 ▪ Revised: November 29, 2018 ▪ Accepted: December 18, 2018

Abstract: The purpose of study is to examine the reliability of analytical rubrics and checklists developed for the assessment of story writing skills by means of generalizability theory. The study group consisted of 52 students attending the 5th grade at primary school and 20 raters in Mersin University. The G study was carried out with the fully crossed hxpxg (story x rater x performance task) design, where the scoring keys were determined as fix facet. Decision Study was carried out by changing the task facet conditions. As a result, it was observed in both scoring keys that the sources of variance related to the stories had a high variance percentage in the main effects while "hp (story and rater interaction effects)" a high variance percentage in the interaction effects. The highest variance in the design belongs to the interaction effect "hpg (story, rater and performance task interaction effects)". This can be an indicator for the existence of different sources of variability and error, which are not included in the design. Examining the G and phi coefficients calculated for both scoring keys, it was determined that scoring with analytic rubrics is more reliable and generalizable. According to the decision studies, it was decided that the number of tasks used in this study is to be most appropriate.

Keywords: *Story writing skills, performance assessment, checklist, rubric, generalizability theory.*

To cite this article: Uzun, N. B., Alici, D., & Aktas, M. (2019). Reliability of the analytic rubric and checklist for the assessment of story writing skills: g and decision study in generalizability theory. *European Journal of Educational Research, 8*(4), 169-180. doi: 10.12973/eu-jer.8.1.169

Introduction

There are four basic skills in language teaching: listening, reading, speaking and writing. It is important for an individual to effectively use of a mother tongue and/or another language in meeting their needs in daily and professional lives; in other words, it is of significance that each of these skills is developed. In this sense, ensuring that students acquire basic language skills is among the most important goals of language education.

According to National Commission on Writing in America's Schools and Colleges (2003, p. 13), writing, one of the basic skills of language teaching, is defined as "a complex intellectual activity that requires students to stretch their minds, improve analytical skills and make a correct and valid distinction". Writing skills are more complicated than other language skills because they require presenting the ideas in an organized and planned manner. According to Hyland (2003), performance in language development depends on the development of writing skills. The development of writing skills is important in educational processes because it is essential for the academic success of students (Kellogg, 2008; Javed, Juan and Nazli, 2013).

Learning about students' writing skills in education plays an important role in the planning and management of educational processes, and effective development of writing skills of individuals. For this reason, assessing the writing skills also get important as the assessment results of the individual's writing skills are taken into account in educational, professional and administrational decisions. However, the assessment of writing is seen as a problematic area for educational researchers and specialists.

Attempts to discuss the question of reliability in assessing writing skills have brought out the use of two basic methods. The first one is the "direct assessment" method in which the writing skills of individuals-students are observed through a composition-type writing example and scored by different raters. The other one is "indirect assessment" method in which writing skills are often observed through objective tests such as multiple choice tests (Stiggins, 1982; Breland, 1983; Cooper, 1984).

* **Corresponding author:**
N. Bilge Uzun , Mersin University, Faculty of Education Department of Educational Sciences, Turkey.
✉ n.bilgeuzun@gmail.com

Many researchers argue that assessing writing skills of individuals through essays, or, in other words, direct assessment is the most effective method. Indirect assessment methods are criticized for not being able to measure writing skills because students do not do “writing” during this assessment (Breland, 1983; Cooper, 1984; Real and Hudson, 1983). According to Coffman (1971, p. 273), "the only way to evaluate a student’s degree of success in a field is to ask him/her questions or problems and to see what performance s/he shows. Writing essays is a direct measure of success as it forms a scientific performance example (cited in Breland, 1983). In his study that compares the deficiencies and superiorities of direct and indirect methods in assessing writing performance, Stiggins (1982) states that direct assessment is more favourable in that it covers more information about the actual writing competence of the individual, the question and answer show high suitability, the exercises can be adapted to various writing conditions related to the real world, writing samples are of high validity, and cost of test development is relatively low. The most important disadvantage of the direct method is the scoring cost.

Direct assessment of writing skills is complex and difficult because there are many sources in this process that may affect the variability of scores. One of the most important sources of variability that can arise in assessing writing performance is "rater". According to Speck and Jones (1998, p. 17), "there are more problems than the solution, and these problems are associated with the inter-rater reliability, the consistency of a single rater, and our absolute responsibility for the grading". One of the main problems in the assessment of writing skills is the reliability of the assessment; in other words, the scoring/scorer/rater reliability. Popham (1990) states that inter- and intra-rater variability sources contributed to the error of measuring in assessment of students' writing skills, which threatens the equity in the writing assessment.

The score that a student gets from writing performance may vary according to the rater and scoring criteria (Breland, 1983). This source of variability can be divided into two sources: "inter-raters" and "intra-rater". There are studies revealing that raters are influenced by such factors as personal identifying information of the answerer (name-surname, gender, ethnic origin, race, etc.), the presentation of the writing task, the linguistic quality of the writing, the scoring sequence of the paper, the rigidity of rater, the place where scoring is done, the scoring method, the scoring criteria, the tendency of the rater to move to the average, the length of the text, and the quality of the text (Branthwaite, Trueman, and Berrisford, 1981; Hamp-Lyons, 1991; Speck and Jones, 1998; Kondo-Brown, 2002; Brown, 2010; Gugiu, Gugiu and Baldus, 2012; Han and Ege, 2013; Wing, Stager and Patil, 2017). All these factors cause the reliability of scoring for writing skills to decrease.

Although there are studies showing that inter-rater reliability plays an important role in the assessment of writing skills (Shohamy, Gordon and Kraemer, 1992), rater reliability does not provide information on the whole of the variability in writing performance. Cooper (1984) describes the sources of variability that could affect the reliability of the exams about writing performance as writer, topic, discourse mode, time limit, appearance, test environment, rater inconsistencies, writing context (content) and sampling error. In the assessment of writing skills, these sources may have effect on the variability of scores separately or through the interaction effect of their different combinations.

One way to reduce variability that can arise from such factors and improve scoring reliability in assessing writing performance is to benefit from scoring keys. Scoring keys are scoring tables developed for teachers or others evaluators to assess the performance of students (Brookhart, 1999). The use of scoring keys in performance assessment enables the scoring to be done in an objective and fair manner and the scoring process to be carried out more effectively. Although checklist, analytic or holistic rubrics are most likely to be seen in the literature, there are also different scoring keys such as focused holistic scoring, atomistic scoring, primary trait scoring, syntactic scoring, and computerized (automatic) scoring (Yamamato, Umemura and Kwano, 2017; Tedick, 2002; Petersen, 1999; Scott and Virginia, 1996; Hamp-Lyons, 1991; Sheila and Brutten, 1990; Breland, 1983; Lloyd-Jones, 1977).

Checklists are lists that are created to show whether certain behaviors, features, or activities exist. Checklists are of the frequently used tools as they are practical and low-cost method, are easy to make marking and produce consistent results. On the other hand, rubrics are assessment tools that contain detailed explanations of each dimension of the trait to be measured. In other words, they help define the explanations regarding the levels of performance desired to be observed. Both tools allow the writing skill to be evaluated in a consistent and reliable way.

There are a number of studies that show the importance of rubrics in assessing writing performance (East, 2009; Rezaei and Lovorn, 2010; Janssen, Meier and Trace, 2015; Ene and Kosobucki, 2016; Fraile, Panadero and Pardo, 2017). Jonsson and Svingby (2007) revealed the benefits of using rubrics in terms of scoring consistency and their competency in facilitating scoring of complex skills, by examining 75 studies conducted in the field. The results of the study suggest that use of rubrics particularly in performance assessments increases the reliability of assessments, assessments can be made with a more comprehensive validity if the validity of the rubrics is ensured, rubrics have a potential for promoting learning and improving teaching as they clearly show the expectations and criteria that facilitate feedback and self-assessment.

Ferrara (1993) states that the generalizability studies conducted in written and other performance assessments are remarkably successful. Generalizability Theory is considered to be a very appropriate approach as it allows to deal with different 'measurement error' sources in these assessments and writing assessments are multifaceted (Brennan, 2001). The theory has an important place in terms of producing information about reliability and validity, considering the errors from various sources of variation.

Generalizability Theory was proposed by Cronbach et al. (1972). A more detailed analysis of measurement errors is provided in the proposed Theory of Expansion (Cronbach, Glasser, Nanda and Rajaratnam, 1972; Brennan, 1992) as an alternative to the classical test theory in which the observed scores are explained based on a single source of error. Generalizability Theory and its corresponding (Generability (G) and Decision (D) studies) aim to separate the error into different sources of variability, or components of variance. The G study focuses on predicting the relative size of these variance components while the D study examines what changes can be made to minimize the error variance on certain surfaces.

Shavelson, Webb, and Rowley (1989) point out that Generalizability Theory can provide a more flexible measurement theory for researchers and is appropriate for practical applications. Generalizability Theory ensures significant advantages in comparison with Classical Test Theory, especially when there are multiple measurement errors in data with complex experimental design (Tobar, Stegner and Kane, 1999: 142-143). While the classical test theory focuses on the fact that each observation or test score has a single true score and a single reliability coefficient (Nunnaly and Bernstein, 1994 cited in: Matt, 2003), G theory also suggests how to deal with the traditional differentiation between reliability and validity. Using the terms "dependability" and "generalizability" instead of "reliability" proves that a unifying reliability and validity is of primary concern (Matt, 2003).

In summary, the reliability problems experienced in the evaluation of writing skills arise from the necessity of the scores of the writing performers by the raters. The objectivity of the evaluation decreases reliability. One of the suggestions to reduce this objectivity is the use of scoring keys. The results of the research indicate that the use of rubric in the performance evaluations increased the reliability of the evaluations. In addition, the use of G Theory is recommended in order to reveal errors from different sources when evaluating performance. This study is designed to search for answers to relevant questions so that writing skills can be measured in a reliable way, to see the resulting sources of error, to make the necessary arrangements and to determine which measuring tool is more reliable to use.

In performance measurements, the results of reliability of the raters are generally given on a single measuring instrument. However, other variability sources and interactions that affect the reliability of the measurement results should be considered. In this respect, it is considered that it is important to make comprehensive reliability analyzes with G theory in the measurement tools that are frequently used in performance measurements and that the study will contribute to the literature with this method.

Methodology

Research Goal

The aim of this study is to examine the reliability of analytic rubrics and checklists developed for assessing story writing skills by means of generalizability theory and to conduct decision studies by changing the condition numbers of some facets. Answers to the following questions were searched:

1) What are the predicted components of variance of the story, the rater, the task and their interaction effects for the different scoring keys?

2) What are the reliability coefficients obtained by scoring the story writing skills in line with the different scoring keys?

3) What is the effect of changing the number of tasks on the g and phi coefficients?

Scoring reliability is important in assessment of writing performance. It is expected with this research that providing reliability of the different scoring keys developed for assessing the story writing skills of children at primary school level, by means of generalizability theory, and also the implementation of decision studies, will significantly contribute to the literature. Moreover, no study appears in the field of literature examining the use of checklists to evaluate the writing performance and investigating the reliability of the tests through generalizability theory. This study is expected to fill this gap and set an example for educators who want to evaluate writing performance.

Type of research

This is a fundamental study as it aimed to determine the reliability of the performance scores obtained from the analytic rubrics and the checklist developed to measure story writing skills of fifth grade students in primary school.

Sample and Data Collection

The study group consists of 52 students studying in the 5th grade of primary school. 20 raters from the teacher candidates attending Mersin University, Departments of Primary School Teacher Education Training and Turkish Teaching were employed to score the stories of the students through checklist and analytic rubric. When the rater sample of the study was formed, the Turkish language teaching department was selected by making purposive sampling when they could have more mastery of the story writing skill. After that, 20 volunteer raters were determined by easily found sampling method. 52 students in the study group were found to be easily found using the sampling method.

The checklist and analytic rubric developed by Aktas (2013) were used in assessing the story writing skills of the students. The checklist and analytic rubric consist of two main criteria as content and format, and four subscales as grammar and spelling, page layout, wording and editing, and a total of 23 behaviors. Each behavior was scored on the checklist (Appendix A) by 1 and 0, depending on whether the students demonstrated the relevant skills, while they were scored as 0-1-2 on the analytic rubric (Appendix B) according to the demonstration level of the relevant behavior.

The 5th grade students of primary school were shown a picture (Appendix C) and were asked to write a story by looking at this picture. Each rater assessed the 52 stories in 10-15 days intervals first with checklist and then analytic rubric.

Analyzing of Data

The data were analyzed on the basis of generalizability theory in this study. Mixed and fully crossed designs were used. Mixed designs in generalizability theory are used when at least one of the facets in the study is fixed. Depending on his/her purpose, the researcher can determine to be generalized facet conditions as fixed or random (Brennan, 2001; Guler, Uyanik and Teker, 2012). If the facet conditions are taken fixed, the results obtained cannot be generalized for other conditions in the universe (Brennan, 2001). Two scoring keys were used in this study: - since scoring keys were selected purposefully and represented certain conditions - a) fixed facet as any generalization was not considered beyond the conditions within the scope of the study, b) random facet for the conditions belonging to other facets (story: h, rater: p, performance tasks: g).

In the G study, components of variance were examined; G and phi coefficients were calculated. In addition, Decision studies were carried out by changing the number of tasks to 10, 30, 40, 50 and 60 in order to determine the appropriate number of tasks for the story writing skills in the scoring keys. In order to answer the first and second sub-problems, different scoring keys' component of variances and reliability coefficients were analyzed by using G study analysis by using Edu G program. In order to answer the third sub-problem, D study analysis was performed by using Edu G program.

Findings / Results

Table 1 (Checklist and Analytic rubric) shows the components of variance predicted as a result of the G study with the fully crossed hxpxg (all stories (h) scored by all raters (p) for all tasks (g)) design in which the scoring keys for the assessment of 52 stories by the 20 raters were determined as fixed facet.

Table 1. The estimated components of variance for G study checklist and analytical rubrics

	Checklist				**Analytic Rubric**			
Source of Variance	**Sum of Squares**	**Mean of Squares**	**Corrected Components**	**%**	**Sum of Squares**	**Mean of Squares**	**Corrected Components**	**%**
h	158,92	3,12	0,0042	**1,7**	539,38	10,58	0,0165	**3,5**
p	10,01	0,53	-0,0004	**0**	37,23	1,96	-0,0007	**0**
g	4,94	0,23	-0,0002	**0**	6,26	0,28	-0,0002	**0**
hp	935,72	0,97	0,0335	**13,6**	2740,33	2,83	0,1081	**22,8**
hg	456,56	0,41	0,0106	**4,3**	553,51	0,49	0,0076	**1,6**
pg	105,32	0,25	0,0011	**0,4**	138,82	0,33	-0,0002	**0**
hpg	4177,7	0,2	0,196	**79,9**	7296,62	0,34	0,3423	**72,1**
Total	**5849,17**			**100**	**11312,2**			**100**

There are 7 variance sources of the fully crossed hxpxg design. According to Table 1, in the design where fixed facet is the checklist, the variance of the story variance source accounts for 1.7% of the total variance and 3.5% of the analytic rubric. This shows the variability between different stories written by different individuals. When the story, rater and task as the main effects are taken into consideration, "different stories written by different individuals" which is the object of the measurement in both scoring keys has the highest component of variance. This is an expected and desired situation when different sources of variance are discussed because the main variability in the generalizability studies is expected to be derived from the basic object of measurement. On the other hand, the fact that variance sources of the other facets are relatively low is interpreted as the rarity of systematic errors originating from sources of variability.

The percentage of variance explained for p and g main effects is seen to be “0” for the checklist and analytic rubric. The finding can be interpreted as the scoring of the raters is consistent and the determined criteria (tasks) for the behaviors to be measured are well defined.

For the checklist and analytic rubric, the “hp” component seems to be the highest variance component in the interaction effects while it is the second largest variance explained in all the variance components. The “hp” variance source accounts for 13.6% of the total variance for the checklist and 22.8% for the analytic rubric. Therefore, the difference between the different raters of the same stories can be said to be high. Another reason may be that the raters experience the halo effect while reading the story.

The “hg” variance source accounts for 4.3% of the total variance for the checklist and 1.6% for the analytic rubric. This finding can be interpreted as that some individuals have higher story writing skills or experiences than other individuals.

The “pg” variance source accounts for 0.4% of the total variance for the checklist and 0% for the analytic rubric. The raters can be said to score the tasks consistently. The variance components of these variance sources in the main effects verify this finding.

The “hpg” variance source is seen to have the highest contribution to the total variance with 79,9% for the checklist and 72.1% for the analytic rubric. This finding shows that the story, task and rater interaction effect and the random error are high.

When the G and Phi coefficients are examined, it is seen that both are 0.62 for the checklist and that the coefficient obtained by using the checklist is low. The G and Phi coefficients are seen to be 0.62 for the analytic rubric and the coefficient obtained from the analytic rubric is higher than of the checklist and is within acceptable limits.

Following the generalizability studies, the findings on the decision studies conducted for both scoring keys are given in Table 2 below.

Table 2. Decision study on the scoring done using checklist and analytic rubric

		Number of tasks (G)					
Checklist		10	23*	30	40	50	60
	G Coefficient	0,53	**0,62**	0,64	0,66	0,67	0,68
	Phi Coefficient	0,53	**0,62**	0,64	0,66	0,67	0,68
Analytic Rubric		Number of tasks (G)					
		10	23*	30	40	50	60
	G Coefficient	0,68	**0,72**	0,73	0,73	0,74	0,74
	Phi Coefficient	0,68	**0,72**	0,73	0,73	0,74	0,74

* The number of tasks in the research.

When Table 2 is examined, it is found that increasing the number of tasks for both scoring keys results in slight increases in the G and Phi coefficients. This is an indicator of the sufficiency of the number of tasks for the measured feature.

Discussion and Conclusion

Since the sequencing of the percentages of both the checklist and the analytic rubric for the total variance is same, the comments below are valid for both scoring keys.

The total variance of the variance component of the stories is high compared with the other main effects, which may be due to the fact that the students writing the story differ in their ability to write stories. On the other hand, the percentage of the other main effects for the total variance is “0”, which can be interpreted that there is no variance arising from the raters and the determined criteria. According to Barbara and Leydens (2000), scoring keys should help ensure consistent scoring regardless of who the raters are. In this sense, the first question that comes to mind when assessing the comprehension of the scoring keys is the question "Are the scoring categories (criteria) well defined?" The criteria included in the developed scoring keys are quite clear. In other words, it can be interpreted that the criteria that reveal the original characteristics of the students regarding their writing skills are included in the scoring keys. The Decision studies supported this interpretation. When the Decision studies are examined, it is concluded that 23 tasks are sufficient and appropriate because the effect of the changing number of tasks on the G and Phi coefficients is relatively low.

The variance component predicted for the "hp" interaction effect shows the incoherence of the raters in assessing the stories (Shavelson and Webb, 1991). The variance component of the "hp" interaction effect predicted in the study is high, which can be interpreted that the difference between different raters' assessment of the same story is high. This does not support the variance component predicted for the rater main effect. However, it can be said, when the variance of the main effects is examined, that the basic effect of this variance is derived from the stories which are the main object of the measurement. In addition, one of the possible reasons for this situation is thought to be the association between raters' past experiences and the stories they read. According to Cooper (1984), papers organized in a bright, interesting or logical way encourage raters to minimize or ignore the mistakes in spelling, technique, usage, and even sentence structure. Another probable reason may be that what Cooper argued is reflected in the raters in different ways.

Even though the variance amount of rater (p) main effect, the fact that there is a significant amount of variance for the interaction effect is because the raters' judgments about the stories are not at the same level of stability for each story (Brennan, 2001; Shavelson and Webb, 1991; Kara and Kelecioglu, 2015).

Although no variance change was observed for the interrater's main effect and the "pg" interaction effect, it can be concluded that the way in which the story was built as a whole by the student (the language, the words, the narration, etc. used by the individuals writing the stories) influences the rater. This interpretation is supported by the fact that the variance percentage of the "pg" interaction effect is very low.

The percentage of the "hg" interaction effect for variance explanations is in the third place. There are differences in the level of fulfilment of the criteria in the scoring key by the individuals who write stories. This difference is supposed to arise from the individuals writing the stories that are the objects of the measurement, not from the task main effect.

The "hpg" residual variance is the source of variance that makes the highest contribution to the total variance in both scoring keys. This may be due to various sources of variance that are not included in the design, as well as the interaction effect. It is thought that the amount of error in the source of variance was increased by the differences arising from raters' residential areas, ages and departments as well as their biased behaviors towards the stories that they read, the errors caused by the environment in which the scoring was made, the fatigue effect due to the long period in scoring the 52 stories, loss in vigilance, reading experiences, and the text type. Indeed, the studies on the impact of scoring time on scoring put the emphasis on fatigue and reading from boredom. Coffman and Kurfman (1968) stated that the longer the reading period and the longer the spread, the lower the tendency to score; accordingly, the first read points are more advantageous. Braddock et al. (1963) emphasized that the fatigue of the raters in the scoring process and their squeezing of the scoring will cause them to score sharper, more tolerant or more unbalanced. The tired and squeezed grader begins to pay more attention to the grammar and technical characteristics of the manuscript during scoring, but ignores the aspects of explanation or composition (cited in Cooper, 1984). McColly (1970) states that the tired rater is much more affected by personal feelings.

When the G and Phi coefficients are examined, it can be said that the scoring made by using analytic rubrics are more reliable and generalizable. There are many studies in the literature that support this finding (Jonsson and Svingby, 2007; East, 2009; Rezaei and Lovorn, 2010; Janssen, Meier and Trace, 2015; Ene and Kosobucki, 2016; Fraile, Panadero and Pardo, 2017).

Since the source of variance in interaction effect and error is high, it can be suggested that the researchers should reduce the probable sources of error or include them in the design. Suggestions to reduce the sources of errors during the implementation can be listed as to give training to the raters by using rubrics and checklists on sample texts and evaluation forms about what they need to pay attention (task description) to during the process they are grading, to give frequent breaks for the raters, to set up the rules about non-subject papers, to create a system to process papers that are original or emotionally disturbing the reader, and so on. Besides, for the similar studies, such additional sources of variance as departments of the raters, book types preferred in their reading, their reading experiences and habits, text type, etc. can be included in the design.

Since the results obtained when the surface conditions are fixed, the results cannot be generalized for other conditions in the universe (Brennan, 2001). The results of this study are limited to checklist and analytic rubric used in the study.

The findings of this study are limited to the characteristics of the scoring keys used. The analytical rubric used in the study is a tool developed in three grades. A generalizability study that can be carried out if the tool is arranged in four or five degrees will provide new findings on how change in the sources of variability will be. Another limitation of the study is related to the type of text that is scored. In this study, generalizability study was conducted by taking into account the story writing skill. In future studies, it may be suggested to change the type of text in the evaluation of the criterion reliability with the generalizability theory.

References

Aktas, M. (2013). *An Investigation of the Reliability of the Scores Obtained Through Rating the Same Performance Task with Three Different Techniques by Different Numbers of Raters According to Generalizability Theory* (Unpublished master's thesis). Mersin University/Institute of Education Sciences, Mersin, Turkey.

Atilgan, H. (2004). *A Research on The Comparability of Generalizability Theory and Multivariate Rasch Model* (Unpublished doctorate thesis). Hacettepe University/ Institute of Social Sciences, Ankara, Turkey.

Bachman, L. F., Lynch, B. K. & Mason, M. (1995). Investigating Variability in Asks and Rater Judgements in a Performance Test of Foreign Language Speaking. *Language Testing, 12*, 238-257.

Branthwaite, A., Trueman, M., & Berrisford, T. (1981). Unreliability of Marking: Further Evidence and a Possible Explanation. *Education Review*, *33*(1), 41-46.

Breland, H. M. (1983). *The Direct Assessment of Writing Skill: A Measurement Review*, *College Board Report No. 83-6, ETS RR No. 83-32*, New York: College Examination Board.

Brennan, R. L. (1992). *Elements of Generalizability Theory (rev. ed.).* Iowa City IA: ACT.

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

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: the Missing Part of Pedagogy. Ashe-Eric Higher Education Report (Vol. 27, No.1).* Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Brown, G. T. (2010). The Validity of Examination Essays in Higher Education: Issues and Responses. *Higher Education Quarterly*, *64*(3), 276-291.

Coffman, W.E. (1971). *Essay Examinations. Educational Measurement (2nd ed.)* R.L. Thorndike ed. Washington D.C.: American Council on Education.

Coffman, W.E. & Kurfman, D.A. (1968). A comparison of two methods of reading essay examinations. *American Educational Research Journal*, 1, 99-107.

Cooper, R. G. (1984). The Strategy-Performance Link in Product Innovation. *R&D Management*, 14: 247–259. doi:10.1111/j.1467-9310.1984.tb00521.

Cooper, P.L. (1984). The Assessment of Writing Ability: A Review of Research, New Jersey: *GRE Board Research Report* No: 82-15R

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley.

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of Generalizability: A Liberation of Reliability Theory. *The British Journal of Statistical Psychology*, 16, 137-163.

East, M. (2009). Assessing the Reliability of a Detailed Analytic Scoring Rubric for Foreign Language Writing. *Assessing Writing,* 14, 88–115.

Ene, E. & Kosobucki, V. (2016). Rubrics and Corrective Feedback in ESL Writing: A Longitudinal Case Study of an L2 Writer. *Assessing Writing*, 30, 3–20.

Ferrara, S. (1993). *Generalizability theory and scaling:* Their roles in writing assessment and implications for performance assessments in other content areas. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta in G. Phillips (Moderator), After a Decade of Authentic Writing Assessment, What Advice Do Frontier States Have to Offer Authentic Assessment Developers in Other Subject Areas?

Guigu, M.R., Guigu, P.C. & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of the grades assigned to undergraduate research papers. *Journal of MultiDiciplinary Assessment*, *8*(19), 26-40.

Guler, N., Uyanik, K. G., & Teker, T. G. (2012). *Genellenebilirlik Kurami.* Ankara: PegemA Yayincilik..

Hamp-Lyons, L. (1991). Reconstructing "Academic Writing Proficiency". In: L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 127–154). Norwood NJ: Ablex.

Han, T., & Ege, I. (2013). Using Generalizability Theory to Examine Classroom Instructors' Analytic Assessment of EFL Writing. *International Journal of Education*, *5*(3), 20-35.

Hyland, K. (2003). *Second Language Writing.* Cambridge: Cambridge University Press.

Janssen, G., Meier, V., & Trace, J. (2015). Building a Better Rubric: Mixed Methods Rubric Revision. *Assessing Writing*, 26, 51–66.

Javed, M., Juan, W. X., & Nazli, S. (2013). A Study of Students' Assessment in Writing Skills of the English Language. *International Journal of Instruction*, *6*(2), 1308-1470.

Jonsson, A., & Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review*, 2, 130–144.

Fraile, J., Panadero, E., & Pardo, R. (2017). Co-Creating Rubrics: The Effects on Self-Regulated Learning, Self-Efficacy and Performance of Establishing Assessment Criteria with Students. *Studies in Educational Assessment*, 53, 69–76.

Kara, Y., & Kelecioglu, H. (2015). Investigation the Effects of the Raters' Qualifications on Determining Cutoff Scores with Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1), 58-71.

Kellogg, R. T. (2008). Training Writing Skills: A Cognitive Developmental Perspective. *Journal of Writing Research*, *1*(1), 1-26.

Kondo-Brown, K. (2002). A Facets Analysis of Rater Bias in Measuring Japanese L2 Writing Performance. *Language Testing*, 19 (1), 3–31.

Lee, Y.-W. , Kantor, R., & Mollaun, P. 2002: *Score Reliability as Essential Prerequisite for Validating New Writing and Speaking Tasks for TOEFL.* Paper presented at the annual meeting of Teachers of English to the Speakers of Other Languages (TESOL). Salt Lake City, UT, April 2002.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), *Evaluating Writing: Describing, Measuring, Judging* (pp. 33-66). Urbana, IL: National Council of Teachers of English.

Matt, G. (2003, November 10). *Generalizability Theory*, Retrieved from http://www.psychology.sdsu.edu/faculty/matt/Pubs/GThtml/GTheory_GEMatt.html

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research*, 64, 148-156.

Moskal, Barbara M. & JonA. Leydens (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Assessment*, 7(10). Available online: http://PAREonline.net/getvn.asp?v=7&n=10

National Commission on Writing in America's Schools and Colleges. (2003). *The Neglected "R": The Need for a Writing Revolution.* New York, NY: College Board.

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

Veal, L.R., & Hudson, S.A. (1983). Direct and Indirect Measures for Large-Scale Assessment of Writing. *Research in the Teaching of English*, 17(3), 290-296.

Petersen, W. (1999). *50 French Oral Communication Activities with Mini-Rubrics*. Auburn Hills, MI: Teacher's Discovery.

Popham, W. J. (1990). *Modern Educational Measurement: A Practitioner's Perspective (2nd ed.).* Englewood Cliffs, NJ: Prentice-Hall.

Rezaei, A.R. & Lovorn, M. (2010). Reliability and Validity of Rubrics for Assessment Through Writing. *Assessing Writing*, 15, 18–39.

Scott, Virgina M. (1996). *Rethinking Foreign Language Writing.* Boston, MA: Heinle & Heinle.

Shavelson, R. J, & Webb, N. M. (1981). Generalizability Theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.

Shavelson, R. J, Rowley, G.L. & Webb, N. M. (1989). Generalizability Theory. *American Psychology*, 44, 922-932.

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

Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, 76, 27-33. http://dx.doi.org/10.1111/j.1540-4781.1992.tb02574.x

Speck, B. W., & Jones, T. R. (1998). Direction in the Grading of Writing. In F. Zak & C. Weaver (Eds.), *The theory and practice of grading writing* (pp. 17-29). Albany, NY State University.

Stiggins, R.J. (1982). A Comparison of Direct and Indirect Writing Assessment Methods. *Research in the Teaching of English*, 16(2), 101-114.

Struthers, L., Lapadat, J. C. & MacMillan, P.D. (2013). Assessing Cohesion in Children's Writing: Development of a Checklist. *Assessing Writing*, 18, 187–201.

Tedick, D. J. (2002). *Proficiency-oriented language instruction and assessment: Standards, philosophies, and considerations for assessment.* In Minnesota Articulation Project, D. J. Tedick (Ed.), Proficiency-oriented language instruction and assessment: A curriculum handbook for teachers (Rev Ed.). CARLA Working Paper Series. Minneapolis, MN: University of Minnesota, The Center for Advanced Research on Language Acquisition.

Tobar, D.A., Stegner, J. & Kane M.T. (1999). The Use of Generalizability Theory in Examining the Dependability of Scores on the Profile of Mood States. *Measurement In Physical Education And Exercise Science*, *3*(3), 141-156.

Wind, S.A., Stager, C., & Patil, Y.J. (2017). Exploring the Relationship Between Textual Characteristics and Rating Quality in Rater-Mediated Writing Assessments: An Illustration with L1 and L2 Writing Assessments. *Assessing Writing*, 34, 1-15.

Yamamoto M., Umemura N., & Kawano H. (2018) *Automated Essay Scoring System Based on Rubric*. In: Lee R. (eds) Applied Computing & Information Technology. ACIT 2017. Studies in Computational Intelligence, vol 727. Springer, Cham.

Appendix A

Ana Olcutler	Alt Olcutler	Alt Olcutlerin Davranis Gostergeleri	EVET	HAYIR
BICIM	DILBILGISI VE IMLA	Tamlama, deyim gibi kaliplar yerinde ve dogru kullanilmistir.		
		Butun kelimeler dogru yazilmistir.		
		Butun kelimeler yerinde kullanilmistir.		
		Noktalama isaretlerinin kullanimi yerinde ve dogrudur.		
	SAYFA DUZENI	Sayfa kenarlarinda bosluk birakilmistir.		
		Yazi tamamen okunaklidir.		
ICERIK	ANLATIM	Cumle kuruluslari tamamen dogrudur.		
		Hikâyeye baslik verilmistir ve icerikle uyumludur.		
		Kullanilan dil akicidir.		
		Hikâyenin konusu acik ve anlasilirdir.		
		Cumleler arasindaki gecisler uygundur.		
		Verilen ornekler yerinde ve yeterlidir.		
		Yapilan betimlemeler yerinde ve uygundur.		
	DUZENLEME	Hikâye kullanilan materyale dayali yazilmistir.		
		Hikâyenin ana fikri acik ve anlasilirdir.		
		Hikâyenin serim, dugum ve cozum bolumleri vardir ve uygundur.		
		Serim, dugum ve cozum bolumleri kendi icinde tutarlidir.		
		Hikâyede anlam bakimindan bir butunluk saglanmistir.		
		Hikâyenin kahramanlari belirgindir.		
		Hikâyenin gectigi yer ve mekân belirgindir.		
		Hikâyenin gectigi zaman ve mevsim belirgindir.		
		Hikâyede gecen olaylarin zaman siralamasina dikkat edilmistir.		
		Hikâyede zaman, kisi, olay orgusu kurulmustur.		

Appendix B

Ana Olcutler	Alt Olcutler	Alt Olcutlerin Davranis Gostergeleri			Puan
		IYI (2 Puan)	ORTA (1 Puan)	KOTU (0 Puan)	
BICIM	DILBILGISI VE IMLA	Tamlama, deyim gibi kaliplar yerinde ve dogru kullanilmistir.	Tamlama, deyim gibi kaliplar kullanilmistir ancak bazilarinin kullanimi yanlistir.	Tamlama, deyim gibi kaliplar yerinde ve dogru kullanilmamistir.	
		Butun kelimeler dogru yazilmistir.	Bazi kelimeler yanlis yazilmistir.	Kelimelerin bircogu yanlis yazilmistir.	
		Butun kelimeler yerinde kullanilmistir.	Bazi kelimeler yerinde kullanilamamistir.	Kelimelerin bircogunu yerinde kullanilamamistir.	
		Noktalama isaretlerinin kullanimi yerinde ve dogrudur.	Noktalama isaretleri kullanilmistir fakat dogru ve yerinde degildir.	Noktalama isaretleri yerinde ve dogru kullanilmamistir.	
	SAYFA DUZENI	Sayfa kenarlarinda bosluk birakilmistir.	Sayfa kenarlarinda bosluk birakilmistir ancak uygun degildir.	Sayfa kenarlarinda bosluk birakilmamistir.	
		Yazi tamamen okunaklidir.	Yazi kismen okunaklidir.	Yazi okunakli degildir.	
ICERIK	ANLATIM	Cumle kuruluslari tamamen dogrudur.	Cumle kuruluslari yeterince dogru degildir.	Cumle kuruluslari dogru degildir.	
		Hikâyeye baslik verilmistir ve icerikle uyumludur.	Hikâyeye baslik verilmistir fakat icerikle yeterince uyumlu degildir.	Hikâyeye hic baslik verilmemistir ya da verilen baslik icerikle hic uyumlu degildir.	
		Kullanilan dil akicidir.	Kullanilan dil yeterince akici degildir.	Kullanilan dil akici degildir.	
		Hikâyenin konusu acik ve anlasilirdir.	Hikâyenin konusu kismen acik ve anlasilirdir.	Hikâyenin konusu acik ve anlasilir degildir.	
		Cumleler arasindaki gecisler uygundur.	Cumleler arasindaki gecisler yeterince uygun degildir.	Cumleler arasindaki gecisler uygun degildir.	
		Verilen ornekler yerinde ve yeterlidir.	Verilen ornekler kismen yerinde ve yeterlidir.	Verilen ornekler yerinde ve yeterli degildir.	
		Yapilan betimlemeler yerinde ve uygundur.	Yapilan betimlemeler yeterince uygun degildir.	Betimleme yapilmamistir ya da yapilan betimlemeler yerinde ve uygun degildir.	
	DUZENLEME	Hikâye kullanilan materyale dayali yazilmistir.	Hikâye kismen kullanilan materyale dayali yazilmistir.	Hikâye kullanilan materyale dayali yazilmamistir.	
		Hikâyenin ana fikri acik ve anlasilirdir.	Hikâyenin ana fikri vardir ancak yeterince acik ve anlasilir degildir.	Hikâyenin bir ana fikri yoktur ya da anlasilmamaktadir.	
		Hikâyenin serim, dugum ve cozum bolumleri vardir ve uygundur.	Hikâyenin serim, dugum ve cozum bolumleri vardir ancak yeterince uygun degildir.	Hikâyenin serim, dugum ve cozum bolumleri yoktur ya da hic uygun degildir.	
		Serim, dugum ve cozum bolumleri kendi icinde tutarlidir.	Serim, dugum ve cozum bolumleri kendi icinde yeterince tutarli degildir.	Serim, dugum ve cozum bolumleri yoktur veya kendi icinde tutarli degildir.	
		Hikâyede anlam bakimindan bir butunluk saglanmistir.	Hikâyede anlam bakimindan yeterince butunluk saglanamamistir.	Hikâyede anlam bakimindan butunluk saglanamamistir.	
		Hikâyenin kahramanlari belirgindir.	Hikâyenin kahramanlarinda kismen bilirsizlik vardir.	Hikâyenin kahramanlari belirsizdir.	
		Hikâyenin gectigi yer ve mekân belirgindir.	Hikâyenin gectigi yer ve mekân yeterince belirgin degildir.	Hikâyenin gectigi yer ve mekân belirgin degildir.	
		Hikâyenin gectigi zaman ve mevsim belirgindir.	Hikâyenin gectigi zaman ve mevsim yeterince belirgin degildir.	Hikâyenin gectigi zaman ve mevsim belirgin degildir.	
		Hikâyede gecen olaylarin zaman siralamasina dikkat edilmistir.	Hikâyede gecen olaylarin zaman bakimindan siralamasinda belirsizlik vardir.	Hikâyede gecen olaylarin zaman siralamasina dikkat edilmemistir.	
		Hikâyede zaman, kisi, olay orgusu kurulmustur.	Hikâyede zaman, kisi, olay orgusu kurulmustur fakat tutarli ve acik degildir.	Hikâyede zaman, kisi, olay orgusu yoktur veya tamamen tutarsizdir.	

Appendix C