

# Using Generalizability Theory to Examine Classroom Instructors' Analytic Evaluation of EFL Writing

Turgay Han<sup>1,\*</sup> & İlhami Ege<sup>2</sup>

<sup>1</sup>Dept. of English Language and Literature, Faculty of Science and Letters, Kafkas University, 36100 Kars, Turkey

<sup>2</sup>Dept. of Classroom Teaching, Faculty of Education, Kafkas University, 36100 Kars, Turkey

\*Corresponding author: Tel: 90-554-622-0962 E-mail: turgayhan@kafkas.edu.tr

Received: May 17, 2013

Accepted: July 4, 2013

Published: July 22, 2013

doi:10.5296/ije.v5i3.3713

URL: <http://dx.doi.org/10.5296/ije.v5i3.3713>

## Abstract

Using G-theory as a theoretical framework, this study was intended to examine the variability and reliability of classroom instructors' analytic assessments of EFL writing by undergraduate students at a Turkish university. Ninety-four EFL papers by Turkish-speaking students in a large-scale classroom-based English proficiency exam were scored analytically by three EFL raters. The results showed great rater variation. Ratings based on two assessment categories (e.g. communicative level and linguistic accuracy level) were also obtained. The variance component for scoring categories (c) did explain total score variance (7.25% of the total variance), suggesting that there was difference in the writing scores that could be attributed to the scoring category itself. Further, the dependability coefficient was .53 for the current scenario and even when the numbers of raters were increased to 10 the dependability of coefficient was .79. This difference had tremendous impact on the reliability of analytic scoring of EFL papers. The findings of this study provide evidence that the classroom teachers should be appropriately trained to score EFL compositions. Important implications are discussed.

**Keywords:** EFL writing assessment; rating variability; rating reliability; generalizability theory

## 1. Introduction

Research with English-as-a-second-language (ESL) and English-as-a-foreign-language (EFL) students has shown that the direct assessment is both complex and challenging (Barkaoui, 2008; Connor-Linton, 1995; Hamp-Lyons, 1991; Huang, 2007, 2008, 2009, 2011; Huang & Foote, 2010; Huang & Han, 2013; Sakyi, 2000). This is because multiple sources contribute to the variability of ESL/EFL students' writing scores. On one end, their age, first language, home culture, style of written communication, English proficiency, and the writing tasks can affect their writing performance to some extent (e.g. Hinkel, 2002; Huang, 2007, 2009, 2011, 2012; Huang & Foote, 2010; Huang & Han, 2013; Kormos, 2011; Kroll, 1990; Shaw & Liu, 1998; Weigle, 2002; Yang, 2001); on the other end, other factors such as essay features, scoring methods, raters' mother tongue, professional background, gender, experience, and type and amount of training can affect rater behavior and outcomes (Barkaoui, 2008; Brown, 1991; Cumming, Kantor & Powers, 2001; Huang, 2007, 2008, 2009, 2011; Huang & Foote, 2010; Huang & Han, 2013; Sakyi, 2000; Shi, 2001; Shohamy, Gordon & Kraemer, 1992; Weigle, 1994, 1999, 2002; Weigle, Boldt, & Valsecchi, 2003).

In the field of EFL/ESL writing assessments, the research has focused on improving consistency and accuracy of the ratings (Connor-Linton, 1995). The variability of the criteria of the raters can be counted as the outstanding source for these inconsistencies as some raters may look for the quality of content and some other may look for the organization (Weigle, 2002) and some others may consider these text feature differently based on proficiency level of essays (Cumming, 1990; Shi, 2001), in this sense, another essay-rater interaction regard raters' varying background such as composition teaching, rating experience, cultural background, training, and expectations and these variables can be very influential in determining scores on writing tasks (Weigle, 2002). This study examines the variability and reliability of EFL writing assessment using generalizability (*G*-) theory rather than classical test theory (CTT) and attempts to extend the knowledge base by examining undergraduate EFL students' of writing samples at a Turkish university.

## 2. Literature Review

Much research in ESL/EFL writing assessments examines scores assigned by markers or raters to investigate validity and reliability of tests and scores, and the evaluations. The CTT approach, the IRT approach (e.g., multi-faceted Rasch measurement), and the *G*-theory approach are the three theoretical frameworks that are used to address variability and reliability issues in the assessment of ESL/EFL writing (Huang, 2007).

### *2.1 The CTT Approach and the G-Theory Approach*

Historically, researchers in the field of second language testing have employed various evaluation techniques (e.g., analysis of variance, regression analysis, factor analysis) to explore testing data (Bolus, Hinofotis, & Bailey, 1982). They have principally used CTT as the theoretical frameworks of their investigations (Brennan, 2001a, 2001b) especially in detection of rater variation in performance assessment situations (Huang, 2007).

CTT is the simplest measurement model and has been widely used to determine reliability of measurements (Bachman, 2004; Eason, 1989). CTT assumes that all measurement errors are random and reliable test scores are a reflection of the test takers' true ability and not the measurement errors (Bachman, 2004). Random measurement errors make a respondent's observed score higher or lower than his or her true score, and therefore lead to unreliable scores (Kieffer, 1998). A true score represents the actual performance of a respondent and is completely reliable; whereas an observed score is given for the generated performance and may not be sufficiently reliable (Kieffer, 1998).

G-theory is a statistical method used to evaluate the dependability of behavioral measurements (Webb & Shavelson, 2005). Dependability refers to the accuracy of generalizing from an individual's observed score on a test to the average score received under all possible conditions (Shavelson & Webb, 1991). Ferrara (1993) states that the G-theory approach

*"...has an important role in all forms of educational assessment, including direct writing assessments and performance assessments in other content areas. More than 20 years ago Coffman (1971) helped set the stage for the use of generalizability analyses of writing assessments when he stated that 'there is a need for studies that control the various sources of error' in writing assessments (p. 282)" (p.2).*

Further, there is a trend towards the use of G-theory in performance assessment, as Eason (1989) states: "...there is every possibility that reflective researchers will increasingly turn to generalizability theory as the measurement model of choice" (p. 21).

The followings are some major strengths of G-theory:

- 1) G-theory estimates multiple sources of variability simultaneously in a single analysis whereas one source of variance separately can be estimated in a single analysis (Shavelson & Webb, 1991).
- 2) G-theory can estimate the magnitude of main and interaction effects of sources of variance (Shavelson & Webb, 1991).
- 3) G-theory enables the calculation of two different reliability coefficients related to decisions based on both the interpretation of the absolute (criterion-referenced) level of scores (Phi coefficient) and of the relative (norm-referenced) level of scores (G coefficient) while CTT enables the calculation of the reliability coefficient for norm-referenced testing situations (Shavelson & Webb, 1991).
- 4) G-theory enables researchers to make decisions about how to reduce the effect of error variance on the true score (Shavelson & Webb, 1991; Güler, 2009; Swartz et al, 1999) whereas CTT can only estimate a single measurement error, such as item, time, rater, form, etc., at a time (Brennan, 2001a, 2001b).
- 5) Alternatively decision- (D-) studies enables researchers to design a measurement protocol to detect the efficiency or cost effectiveness of administering a different number of items or

forms on a different number of occasions (Kieffer, 1998). Therefore, decisions about a person on the basis of his/her test score can be made with minimum error of measurement (Huang, 2007); however, CTT can calculate and forecast the efficiency of a single source of error source (e.g. number of items for maximum reliability) using Sperman-Brown formula (Shavelson & Webb, 1991).

Several empirical studies have recently used G-theory to examine the reliability and validity of EFL/ESL writing scores (Gebril, 2010; Huang, 2008, 2012; Huang & Foote, 2010; Huang & Han, 2013; Han, 2013; Swartz et al, 1999). For example, Swartz et al. (1999) used G-theory to investigate the reliability of holistic and analytic writing scores as well as the influence of raters and the use of writing scores (absolute versus relative decisions) on the reliability of writing scores in either standardized test or classroom-based assessments. The results showed that when the number of raters was reduced and absolute decisions were made, the reliability coefficients for the writing scores declined. These results proved that G-theory is a powerful and flexible approach that allows multiple sources of error variance to be estimated simultaneously in order to determine the reliability of test scores.

In his quantitative study, Huang (2008) used the G- theory approach to examine the rating variability and reliability of scores assigned to ESL essays and Native English (NE) essays in large-scale secondary school writing assessment contexts. A series of generalizability studies and decision studies were conducted to determine differences in score variation between ESL and NE essays. In another study, Huang and Foote (2010) examined score variations and differences between ESL students' papers and NE students' papers in small-scale university classroom assessment context. G-theory was used for data analysis. The results showed that there were differences in consistency and precision between the scores assigned to ESL papers and NE papers. These results raised some concerns about the fairness of ESL writing assessments.

Further, using a multivariate generalizability analysis Gebril (2010) investigated the effects of two different writing tasks (reading-to-write and writing only tasks) and rater facets on composite score generalizability. Data consisted of each of 115 examinees' writings, based on two writing-only and two reading-to-write tasks. The results showed that a composite of the two tasks is as reliable as scores obtained from either writing-only or reading-to-write tasks.

Most recently, Huang (2012) used G- theory to examine the accuracy and validity of the writing scores assigned to ESL students in provincial English examinations. Conducting a series of G-studies and D-studies for three years writing scores obtained in this large-scale exam, if there are any differences between the accuracy and construct validity of the analytic scores assigned to ESL students and to NE students were investigated. The results indicated that there were differences in score accuracy between ESL and NE students. The G-coefficients for ESL students were significantly lower than those for NE students in all three years. Further, there were significantly less convergent validity in one year and less discriminant validity in all three years of the scores assigned to ESL students than to NE students. As a result, this study showed that writing scores assigned to ESL and NE students were significantly different in terms of accuracy and construct validity and these findings

raised a potential question about the presence of bias in the assessment of ESL students' writing.

In 2013, Huang and Han examined the impact of scoring methods on the reliability and variability of EFL writings by undergraduate students at a Turkish university, using G-theory approach. The results showed greater rater variation for holistic scores than for analytic scores of EFL papers. Further, there was a large difference in the G-coefficients between holistic (with a G-coefficient of .64) and analytic scoring (with a G-coefficient of .90) and this difference had tremendous impact on the reliability of holistic scoring of EFL papers. The findings of this study provide evidence that analytic scoring is more appropriate and effective than holistic scoring for professors to score EFL compositions.

In the same year Han (2013) examined the impact of scoring methods and rater training on the classroom-based writing assessment scores, using G-theory as the framework of the study. The results showed that with careful training, holistic scoring could produce comparable consistency and reliability as analytic scoring.

Using G-theory as a framework for analysis, the purpose of this study was to examine classroom instructors' analytic evaluations of EFL students' writing at a Turkish university. Specifically, the following three research questions guided the study:

- 1) Is there any significant difference among the three ratings of the same EFL papers?
- 2) What are sources of score variation contributing to the score variability of the analytic scores assigned to EFL papers?
- 3) What is the reliability of the analytic scores assigned to EFL papers?

### **3. Method**

Using G-theory in this quantitative research, the purpose of this study was to examine classroom instructors' analytic evaluation of EFL writing by undergraduate preparatory class students in a Turkish university. Quantitatively, the rating variability and reliability of English-as-a-foreign-language (EFL) writing scores assigned by three EFL writing course teachers were examined, using SPSS statistics and G-studies. The guiding research question was: "Are there any differences in the rating variability and reliability of EFL students' analytic writing scores?"

#### *3.1 Description of the Data Set*

The Schools of Foreign Languages of a state university in Turkey provided the writing samples necessary for the analyses. Data for this study were collected in two stages. In the first stage, permissions received from the manager office of the School of Foreign Languages of the University. In the second stage, 94 pen-paper-based short compositions written by EFL prep-class students in the writing section of the University English Proficiency Exam that took place in the spring semester of 2011-2012 academic years were selected.

Students who attended English preparatory classes took the proficiency exam in the spring semester in 2012 before starting studying English Language and Literature. A criterion-referenced framework rather than a norm-referenced framework is used in the assessment of writing scores obtained from the exam. The proficiency exam was implemented in three steps in two exam days. First, the students took a grammar and reading comprehension skill test that including 75 multiple-choice questions session in the first exam day. Second, the students attended listening and speaking exam sessions in the morning of the second day where they responded audio-visual stimulus and their responses were rated by course teachers using a rubric. Third, the students attended writing exam session in the afternoon of the second day. In the writing session, they selected one of the tasks proposed in the exam and wrote a short argumentative, descriptive or explanatory composition in 60 minutes. Three writing course teachers rated the compositions using a 6-point analytic rubric. Students need to score 70 out of 100 in order to pass the exam (KAÜ, 2012).

### *3.2 The Writing Samples*

Totally, 94 papers were selected from the proficiency Spring Exam Writing Sections for this study. Each rater then scored these papers analytically independently. The samples were 180-200-word short descriptive compositions on one of the three optional topics given.

### *3.3 The Raters*

The three writing course teachers (two males and one female) who were lecturers with various teaching backgrounds rated the writing samples. One of the male raters and the female rater were doing their PhD studies, and the other male rater was doing his MA study in the field of interdisciplinary EFL and having at least more than one year of experience in EFL teaching and assessment at the time of doing the ratings.

These three raters were all graduates from English Language Teaching departments at different Turkish universities. They had the same L1 background (Turkish) and were all proficient non-native speakers of English. The ages of the three raters ranged from 25 to 30. Their experiences in teaching EFL and scoring EFL essays were different.

### *3.4 The Analytic Rating Scale*

The writing scoring rubric used in rating the writing samples was 6- point analytic scale that was an adapted version of European Portfolio Writing Assessment Scale and it was used to be the department rubric.

The criteria used to evaluate the essays in the 6-point analytic rubric were organized under the following two analytic criteria: a) linguistic structure b) communicative structure. A score between 0-1 represents the lowest writing performance, a score between 1-2 represents the lowest writing performance, and a score between 2-3 represents the highest writing performance of the examinees.

### *3.5 The Rating Procedure*

Each rater scored the descriptive short compositions independently, using the analytic

department rubric. All raters rated papers more than in two or three sessions by referring to scoring rubrics but not comparing the scores they assigned to each paper. The scoring took place at raters' homes and offices to avoid discussions among raters.

### *3.6 The Descriptive Statistics and the G-Theory Analyses*

#### 3.6.1 The Descriptive and Inferential Statistics

Descriptive statistical analysis (the mean and standard deviation) and paired sample t-tests were conducted for the analytic writing scores assigned by the three raters for each paper. The purpose of these statistical analyses was to examine if there was a significant mean score difference among analytic scores assigned by the three raters.

#### 3.6.2 G-studies

Using G-theory framework, data were analyzed in the additional stages: 1) paper-by-rater random effects G-studies, 2) Person-by-category-by-rater random effects G-studies and 3) calculation of G-coefficients (Huang & Foote, 2010).

The paper-by-rater (p x r) random effects G-study was conducted for 94 writing samples. The purpose of these G-studies was to obtain information the analytic scores in terms of score variability and reliability. With the implementation of this G-study, the three independent sources of variation, namely, paper (p), rater (r), and paper-by-rater (p x r) were obtained. G-coefficients were then calculated for examining the reliability, using the obtained variance components.

In the study, three raters scored these 94 papers analytically based on two assessment categories (e.g. communicative level and linguistic accuracy level). This resulted in 94 persons (p) and 188 scores (94 scores for each category), each person receiving three different scores from three raters (r) through communicative scoring level and linguistic accuracy level (category, c). Therefore, this constitutes a fully crossed G-study p x c x r design.

A person-by-category-by-rater (p x c x r) random effects G-study analysis was conducted. The variance component estimates for the seven independent sources of variation: person (p), rater (r), category (c), person-by-rater (p x r), person-by-category (p x c), category-by-rater (c x r), and person-by-rater-by-category (p x r x c) were obtained.

### *3.7 Calculation of G-coefficients*

“Dependability coefficients” are used in a criterion-referenced score interpretation and are denoted by  $\Phi$ . It is the analogue of a reliability coefficient in CTT (Huang, 2007). A dependability coefficient is the ratio of the universe score variance to itself plus absolute error

variance ( $\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}$ ) (Huang, 2007).

Based on the paper-by-rater (p x r) random effects G-study results, G-coefficients were calculated. The purpose of calculating the G-coefficients was to examine the reliability the

scores assigned to EFL papers.

The computer program GENOVA (Crick & Brennan, 1983) was used for the G- and D-studies (Huang, 2012). GENOVA is a computer program used to estimate the variance components for the main and interaction effects and their standard errors where the design is balanced (Huang, 2012). The program also computes the G-coefficients ( $E\rho^2$ ) and dependability coefficients ( $\Phi(\lambda)$ ) for different values of the cut-score  $\lambda$  (cf. Huang, 2007).

## 4. Results

### 4.1 Descriptive Results

Table 1. provides the descriptive statistics for the total scores assigned to the EFL papers by three raters. Both the mean and standard deviation of the ratings shows that the three raters assigned very different scores. Hence the raters could be considered less consistent in using the analytic scale for rating the EFL papers. Comparing the ratings by EFL raters, the results show that Rater #1 assigned the lowest scores for EFL students' performance in writing.

**Table 1:** Descriptive Statistics

	N	Mean	Std. Deviation
R1	94	1.8617	1.81156
R2	94	3.8830	1.17186
R3	94	3.4681	1.34965

Table 2.1.provides the descriptive statistics for the scores assigned to communicative features of the EFL papers by three raters. Both the mean and standard deviation of the ratings shows that the three raters assigned very different scores to this category. Hence the raters could be considered less consistent in using the analytic scale for this rating category. Comparing the ratings by EFL raters, again, Rater #2 assigned the highest scores for EFL students' communicative performance in writing.

**Table 2.1.** Descriptive Statistics for Communication Category

	N	Mean	Std. Deviation
R1	94	1.0532	.95453
R2	94	2.2021	.68124
R3	94	1.9255	.79297

Table 2.2. provides the descriptive statistics for the scores assigned to linguistic features of the EFL papers by three raters. Both the mean and standard deviation of the ratings shows that the three raters assigned very different scores to this category. Again, the raters could be

considered less consistent in rating linguistic features of the writings. Comparing the ratings by EFL raters, again, Rater #2 assigned the highest scores for EFL students' communicative performance in writing.

**Table 2.2.** Descriptive Statistics for Language Use Category

	N	Mean	Std. Deviation
R1	94	.8085	.90728
R2	94	1.6809	.59048
R3	94	1.5426	.72830

#### 4.2 Inferential Statistics

Table 2. shows that there was a significant difference among raters' analytic ratings ( $p < .01$ ). The analytic scores assigned to these EFL papers by the three raters were significantly different from each other. Again, the inferential statistical results confirmed the descriptive statistical results, suggesting that each rater scored the EFL papers very differently.

**Table 2:** Paired Sample Statistics

		Paired Differences					
		Mean	s.d.	Std. Error Mean	t	df	Sig. (2-tailed)*
Pair 1	RATER 1-RATER 2	-2.02128	1.59959	.16498	-12.251	93	.000
Pair 2	RATER 1-RATER 3	-1.60638	1.80333	.18600	-8.636	93	.000
Pair 3	RATER 2 –RATER 3	.41489	1.19506	.12326	3.366	93	.001

\* Note: indicates significant difference at the .01 level

#### 4.3 Person-by-rater Random Effects G-study Results

A person-by-rater ( $p \times r$ ) random effect G-study was conducted for the analytic writing scores. The purpose of this G-study was to obtain information the score variability and reliability. The results are presented in Table 3.

The residual yielded the largest variance component (38.82% of the total variance). The residual contains the variability due to the interaction between raters and papers, and other unexplained systematic and unsystematic sources of error (cf. Huang, 2007, 2008).

For Topic One, as shown in Table 3., person (p), yielded the third largest variance component (26.77 % of the total variance), suggesting that the 36 EFL papers were considerably different in terms of quality.

Rater (r) yielded the second largest variance component (34.41% of the total variance), suggesting that raters did differ considerably from one another in terms of leniency of

marking these EFL papers.

**Table 3:** Variance Components for Random Effects p x r G-study Designs

Source of Variability	df	$\sigma^2$	%
p	93	0.6519	26.77
r	2	0.8383	34.41
pr	186	0.9454	38.82
Total	281	2.4356	100

#### 4.4 Person-by-Category-by-rater Random Effects G-study Results

The person-by-category-by-rater random effects G-study was conducted for the three raters analytic scores. The results are presented in Table 4.

**Table 4:** Variance Components for a Random Effects p x c x r G-Study Design

Source of Variability	df	$\sigma^2$	%
p	94	0.2336	23.82
c	1	0.0711	7.25
r	2	0.2854	29.09
pc	94	0.0008	0.08
pr	188	0.2269	23.13
cr	2	0.0085	0.86
pcr	188	1.1547	15.77
Total	569	0.9810	100

As shown in Table 4., Rater (r) yielded the largest variance component (29.09% of the total variance), indicating that raters did differ from one another in terms of leniency of marking these papers.

Person (p), the object of measurement, yielded the second largest variance component (23.82 % of the total variance), suggesting that the 94 EFL papers were substantially different in terms of quality.

Person-by-rater (pr) yielded the third largest variance component (23.13% of the total variance), indicating that raters marked all papers very differently.

The residual yielded the fourth largest variance component (15.77% of the total variance). The residual contains the variability due to the interaction between raters, scoring categories, persons, and other unexplained systematic and unsystematic sources of error.

The variance component for scoring categories (c) did explain total score variance (7.25% of the total variance), suggesting that there was difference in the writing scores that could be attributed to the scoring category itself.

Category-by-rater (cr) yielded the sixth largest variance component (0.86% of the total variance), indicating there was considerable consistency in terms of rating severity or leniency across scoring methods.

Person-by-category (pc) yielded the seventh largest variance component (only 0.08% of the total variance), indicating that these papers are relatively similar in terms of qualities across scoring categories.

#### *4.5 Calculation of Dependability Coefficients*

Using the formula above and the person-by-rater random effects G-studies variance component results, the dependability coefficients topic were calculated and presented in Table 5. As shown in Table 5., the dependability coefficient obtained for the for the current three-rater scenario was .52. Further, the results show that increasing the number of raters to 10 for the holistic scoring method would result in a dependability coefficient of .79.

**Table 5:** Summary of G-coefficients

Number of Papers	Number of Raters	Dependability coefficients of Analytic Scoring
94	1	.27
94	2	.73
94	3	.52
94	4	.59
94	5	.65
94	6	.69
94	7	.72
94	8	.75
94	9	.77
94	10	.79

## **5. Conclusion and Discussion**

The first research question attempted to determine if there would be any differences among the three raters' ratings of the same EFL paper. The descriptive statistical results showed that the three raters assigned very different scores. Hence the raters could be considered less consistent in using the analytic scale for rating the EFL papers. Further, the ratings were examined to see if there was any difference between the scores assigned to the communication level and the language use level of the same EFL papers. The descriptive results showed that the three raters assigned very different scores to the communicative features and linguistic features of the EFL papers.

The second research question examined the differences in score variation. The results showed that first; raters did differ from one another in terms of leniency of marking these papers

(rater 29.09% of the total variance). Second, the 94 EFL papers were substantially different in terms of quality component (Person the object of measurement, 23.82 % of the total variance). Third, raters marked all papers very differently (Person-by-rater 23.13% of the total variance). Fourth, there was undesired variability due to the interaction between raters, scoring categories, persons, and other unexplained systematic and unsystematic sources of error (15.77% of the total variance). Although these differences may be due to the scoring categories, the analytic scale used and/or other facets that could have attributed to the score variance (e.g., quality of EFL papers) (Han, 2013; Huang & Han, 2013). Fifth, scoring categories (c) did explain total score variance (7.25% of the total variance), suggesting that there was difference in the writing scores that could be attributed to the scoring categories itself. Next, there was considerable consistency in terms of rating severity or leniency across scoring methods (category-by-rater 0.86% of the total variance). Finally, the EFL papers are relatively similar in terms of qualities across scoring categories (person-by-category only 0.08% of the total variance).

The third research question focused on the reliability of the analytic scores. As previously mentioned, the dependability coefficient obtained for the current three-rater scenario was .52. Further, the results show that increasing the number of raters to 10 for the holistic scoring method would result in a dependability coefficient of .79.

The present study was limited in the following two ways. First, writing task (i.e., only one descriptive essay from each student participants was used in the analysis) and paper qualities were not considered in this study. However, research has shown that different writing tasks impact the scoring variability and reliability of ESL/EFL essays (Huang, 2008; Lee & Kantor, 2005). Second, due to the quantitative nature of analyzing and reporting ESL/EFL writing scores, this study only used a quantitative approach. However, the investigation of empirical evidence for rater variation in ESL/EFL writing assessments, as argued by Connor-Linton (1995), should look more closely at the rating process. Think-aloud protocol analysis is popularly used to investigate the thinking processes and criteria used by raters of ESL/EFL compositions because it provides the “richest evidence” about what raters think and do while rating ESL/EFL essays; and therefore, the research on the rating process can address not only many aspects of rating scale validity issues but also a number of fairness issues (Connor-Linton, 1995; Cumming, 1990; Sakyi, 2000; Vaughan, 1991; Weigle, 1994).

Overall, the results of this study indicate that classroom teachers as raters in their analytic scoring of EFL essays result in differences in terms of consistency and precision. If it is a common practice in Turkish universities that either classroom teachers frequently use their inner criteria or use a single scoring rubric without receiving a rater training while rating EFL students' compositions in classroom-based assessments, rater training is essential.

In conclusion, the findings of this study provide evidence that rater training should be applied to the classroom teachers to get more reliable and fairer writing scores (Han, 2013) because rater training possibly has a direct impact on applying the scoring criteria on the rubric reliably and, therefore, it increases the reliability of the interpreting and scoring dimensions of the rubric (Stuhlmann, et al., 1999; Weigle, 1994, 1998).

The implications for the professors should be the establishment of a clear scoring guide, holistic or analytic, and the adherence to such a scoring guide while marking the ESL/EFL compositions (Huang & Foote, 2010). Only through the rater training can reduce the grading inconsistency.

## References

- Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511667350>
- Barkaoui, K. (2008). *Effects of Scoring Method and Rater Experience on ESL Essay Rating Processes and Outcomes*. (Unpublished Doctoral Dissertation). Canada: University of Toronto.
- Bolus, R.E., Hinofotis, F.B., & Bailey, K.M. (1982). An Introduction to Generalizability Theory in Second Language Research. *Language Learning*, 32(2), 245-258. <http://dx.doi.org/10.1111/j.1467-1770.1982.tb00970.x>
- Brennan, R.L. (2001a). *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag. Retrieved March 30, 2013 from <https://www.google.com.tr/search?hl=tr&tbo=p&tbm=bks&q=isbn:0387952829>
- Brennan, R.L. (2001b). *Generalizability theory*. Iowa: ACT Publications. <http://dx.doi.org/10.1007/978-1-4757-3456-0>
- Brown, J.D. (1991). Do English and ESL Faculties Rate Writing Samples Differently? *TESOL Quarterly*, 25(4), 587-603. <http://dx.doi.org/10.2307/3587078>
- Connor-Linton, J. (1995). Looking Behind the Curtain: What Do L2 Composition Ratings Really Mean? *TESOL Quarterly*, 29(4), 762-765. <http://dx.doi.org/10.2307/3588174>
- Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENERALized Analysis of VARIANCESystem* (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.
- Cumming, A. (1990). Expertise in Evaluating Second Language Composition. *Language Testing*, 7(1), 31-51. <http://dx.doi.org/10.1177/026553229000700104>
- Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL Essays and TOEFL 2000 Prototype Tasks: An Investigation into Raters' Decision Making and Development of a Preliminary Analytic Framework (TOEFL Monograph Series, Report No: 22)*. Princeton, NJ: Educational Testing Service.
- Eason, S. (1989). *Why Generalizability Theory Yields Better Results Than Classical Test Theory* [Proceeding]. Paper Presented at the Annual Meeting of the Mid-South Educational Research Association, November 8-10, 1989, Little Rock: AR. (ERIC Document Reproduction Service No. ED 3 14 434).

- Ferrara, S. (1993). *Generalizability Theory and Scaling: Their Roles in Writing Assessment and Implication for Performance in Other Content Areas* [Proceeding]. Paper Presented at the Annual Meeting of the National Council of Measurement in Education. Atlanta: GA. Retrieved on April 20, 2013 from <http://marces.org/mdarch/pdf/M032030.pdf>
- Gebril, A. (2010). Bringing Reading-to-Write and Writing-Only Assessment Tasks Together: A Generalizability Analysis. *Assessing Writing*, 15(2), 100-117. <http://dx.doi.org/10.1016/j.asw.2010.05.002>
- Güler, N. (2009). Genellenebilirlik Kuramıve SPSS ile GENOVA Programlarıyla Hesaplanan G ve K Çalışmalarına İlişkin Sonuçların Karşılaştırılması. *Education and Science*, 34(154), 93-103.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-277). Norwood, NJ: Ablex.
- Han, T. (2013). *The Impact of Rating Methods and Rater Training on the Variability and Reliability of EFL Students' Classroom-Based Writing Assessments in Turkish Universities: An Investigation of Problems and Solutions*. (Unpublished Doctoral Dissertation). Turkey: Atatürk University.
- Hinkel, E. (2002). *Second language writers' text: Linguistics and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Huang, J. (2007). *Examining the Fairness of Rating ESL Students' Writing on Large-Scale Assessments*. (Unpublished Doctoral Dissertation). Canada: Queen's University.
- Huang, J. (2008). How Accurate Are ESL Students' Holistic Writing Scores on Large-Scale Assessments?—A Generalizability Theory Approach. *Assessing Writing*, 13(3), 201-218. <http://dx.doi.org/10.1016/j.asw.2008.10.002>
- Huang, J. (2009). Factors Affecting the Assessment of ESL Students' Writing. *International Journal of Applied Educational Studies*, 5(1), 1-17. <http://dx.doi.org/10.1299/spacee.5.1>
- Huang, J. (2011). Generalizability Theory As Evidence of Concerns about Fairness in Large-Scale ESL Writing Assessments. *TESOL Journal*, 2(4), 423-443. <http://dx.doi.org/10.5054/tj.2011.269751>
- Huang, J. (2012). Using Generalizability Theory to Examine the Accuracy and Validity of Large-Scale ESL Writing Assessment. *Assessing Writing*, 17, 123-139. <http://dx.doi.org/10.1016/j.asw.2011.12.003>
- Huang, J., & Foote, C.J. (2010). Grading Between Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing? *Language Assessment Quarterly*, 7(3), 219-233.
- Huang, J., & Han, T. (2013). Holistic or analytic – A Dilemma for Professors to Score EFL Essays? *Leadership and Policy Quarterly*, 2(1), 1-18.

- KAÜ. (2012). KAÜ SFL English Proficiency Exam. The English Proficiency Exam of Kafkas University (n.d.). Retrieved December on 15, 2012 from [http://www.kafkas.edu.tr/yabanci\\_diller/yonetmelik.pdf](http://www.kafkas.edu.tr/yabanci_diller/yonetmelik.pdf)
- Kieffer, K. M. (1998). *Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate?* [Proceeding]. *Paper Presented at the Annual Meeting of the South Western Psychological Association*. New Orleans, LA: USA.
- Kormos, J. (2011). Task Complexity and Linguistic and Discourse Features of Narrative Writing Performance. *Journal of Second Language Writing*, 20, 148-161. <http://dx.doi.org/10.1016/j.jslw.2011.02.001>
- Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140-154). Cambridge: Cambridge University Press.
- Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL Writing Test Scores: Evaluating Prototype Tasks and Alternative Rating Schemes* (TOEFL Monograph No. MS-31). Princeton, NJ: ETS.
- Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge: Cambridge University Press.
- Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Premier*. Newbury Park, CA: Sage.
- Shaw, P., & Liu, E. T.-K. (1998). What Develops in the Development of Second Language Writing. *Applied Linguistics*, 19(2), 225-254. <http://dx.doi.org/10.1093/applin/19.2.225>
- Shi, L. (2001). Native- and Nonnative-Speaking EFL Teachers' Evaluation of Chinese Students' English Writing. *Language Testing*, 18(3), 303-325.
- 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>
- Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R.K., & Powers, T. (1999). A Generalizability Study of the Effects of Training on Teachers' Abilities To Rate Children's Writing Using A Rubric. *Journal of Reading Psychology*, 20, 107-127. <http://dx.doi.org/10.1080/027027199278439>
- Swartz, C.W., Hooper, S.R., Montgomery, J.W., Wakely, M.B., De Kruif, R.E.L., Reed, M., Brown, T.T., Levine, M.D., & White, K.P. (1999). Using Generalizability Theory to Estimate the Reliability of Writing Scores Derived From Holistic and Analytic Scoring Methods. *Educational and Psychological Measurement*, 59, 492-506. <http://dx.doi.org/10.1177/00131649921970008>
- Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In

- L.Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Norwood, NJ: Ablex.
- Webb, N.M., & Shavelson, R.J. (2005). Generalizability theory: Overview. In B.S. Everitt & D.C. Howell (Eds), *Encyclopedia of statistics in behavioral science* (pp.717-719). Chichester: John Wiley & Sons.
- Weigle, S. C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, 15, 263-87. <http://dx.doi.org/10.1177/026553229801500205>
- Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of Task and Rater Background on the Evaluation of ESL Writing: A Pilot Study. *TESOL Quarterly*, 37(2), 345-354. <http://dx.doi.org/10.2307/3588510>
- Weigle, S.C. (1994). Effects of Training on Raters of ESL Compositions. *Language Testing*, 11, 197-223. <http://dx.doi.org/10.1177/026553229401100206>
- Weigle, S.C. (1999). Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing*, 6, 145-178. [http://dx.doi.org/10.1016/S1075-2935\(00\)00010-6](http://dx.doi.org/10.1016/S1075-2935(00)00010-6)
- Weigle, S.C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511732997>
- Yang, Y. (2001). *Chinese Inference in English Writing: Cultural and Linguistic Differences* (Harvard Graduate School of Education Report No: FL 027 138). Unpublished manuscript. Retrieved April 20, 2011 from <http://www.eric.ed.gov/PDFS/ED461992.pdf>

### **Copyright Disclaimer**

Copyright reserved by the author(s).

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).