PASAA
Volume 53
January - June 2017

Examining the Impact of Scoring Methods on the Institutional EFL Writing Assessment: A Turkish Perspective

Turgay Han
Ordu University, Turkey
Jinyan Huang
Niagara University, USA

Abstract

Using generalizability (*G*-) theory and rater interviews as both quantitative and qualitative approaches, this study examined the impact of scoring methods (i.e., holistic versus analytic scoring) on the scoring variability and reliability of an EFL institutional writing assessment at a Turkish university. Ten raters were invited to rate 36 undergraduate argumentative essays first holistically and then analytically, with a three-week time interval. The quantitative results indicated that with proper rater training holistic scoring can produce as reliable and dependable assessment outcomes as analytic scoring. Similarly, the qualitative results revealed that all raters prefer using the holistic scoring method because it could help them not only assign fair and objective scores to essays but also facilitate their scoring process. Further, most raters agreed that the content of an essay was the most important factor that most affected their holistic scoring decision making of an essay. In contrast, all aspects of an essay (e.g., grammar, content, or organization) jointly affected their analytic scoring decision

making of an essay. Important implications for EFL writing assessment professionals in the institutional assessment context are discussed.

Keywords: EFL writing assessment, scoring methods, generalizability (*G*-) theory, rater interviews, rating variability, rating reliability.

Introduction

Assessing writing is a common type of language performance assessment (Barkaoui, 2008; Connor-Linton, 1995; Huang, 2012). Unlike multiple-choice assessment, the direct assessment of English as a second language (ESL) or English as a foreign language (EFL) students' writing is both complex and challenging (Hamp-Lyons, 1995; Huang, 2010). Not only do the sources such as age, mother tongue, culture, proficiency level, and task type (Han, 2013; Hinkel, 2002; Huang, 2009, 2011, 2012; Huang, 2010; Huang & Han, 2013; Kormos, 2011; Weigle, 2002; Yang, 2001) lead to the variability of ESL/EFL students' writing scores, but also there are other factors causing variability of scores, such as essay features, rating methods, scorers' L1, background, gender, experience (Alharby, 2006; Cumming, Kantor, & Powers, 2002; Goulden, 1994; Huang, 2008, 2012; Knoch, Read, & Randow, 2007; Lim, 2011; Shi, 2001; Weigle, 1994, 1998; Weigle, Boldt, & Valsechi, 2003).

Although EFL/ESL students' writing performance varies naturally, variability caused by raters and tasks are not desired as they lead to measurement error and unreliability regarding the writing scores (Huang, 2012; Huot, 2002; Sudweeks, Reeve, & Bradshaw, 2005). Raters are central to writing performance assessment; and rater training, rater experience, and rater expertise involve a temporal dimension (Lim, 2011). Therefore, the above factors are the sources of weak reliability, validity, and fairness regarding the ESL/EFL writing scores (Barkaoui, 2008; Huang, 2008, 2012; Jonnson & Svingby, 2007; Weigle, 2002).

Reliability, validity and fairness are the three major topics that are broadly debated in the context of performance assessments (Davies, 2010; Kane, 2010; Xi, 2010). Many studies have examined how factors associated with writing tasks and rater behaviors can impact the reliability, validity, and fairness of ESL/EFL writing assessments (Ebel & Frisbie, 1991; Johnson, Penny, & Gordon, 2009; Song & Caruso, 1996; Weigle, 2002). However, limited research has used the generalizability (*G*-) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and rater interview approaches to examine the impact of scoring methods on the ESL/EFL writing assessments. Using *G*-theory and rater interviews as both quantitative and qualitative approaches, this study examined how scoring methods affect the rating variability and reliability of an EFL institutional writing assessment at a Turkish university.

The Impact of Scoring Methods on ESL/EFL Writing Assessment

For decades, holistic and analytic scoring methods have been used in writing assessment practices extensively (Carr, 2000; Jonsson & Svingby, 2007). These two scoring methods have both advantages and disadvantages. For example, holistic scoring has the highest construct validity; and it is recommended as a tool for certification, placement, proficiency, and research testing (Russikoff, 1995; Weigle, 2002). However, holistic scoring can be referred to as threats to reliability for the reason that it can be extremely subjective owing to "bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next" (Perkins, 1983, p. 653). In contrast, analytical scoring schemes can provide "more detailed information about a test taker's performance in different aspect of writing … for this reason preferred over holistic schemes by many specialists" (Weigle, 2002, pp. 114-115). Even though the analytic scoring process can yield higher inter-rater reliability than the holistic rating, it is more time-consuming, more cost-effective, but less unbiased (East, 2009; Perkins, 1983, Weigle, 2002). The literature discusses the ongoing common debate for teachers, administrators, researchers, and assessment specialists in

choosing a more valid and reliable method to assess ESL/EFL writing performance in both classroom and large-scale exams (Barkaoui, 2008; 2010a, 2010b; Carr, 2000; Knoch, 2009; Lee, 2004; Lee, Gentile, & Kantor, 2009). The following is a brief summary of the related literature.

First, in the ESL/EFL writing assessment context, several studies investigated the effectiveness of holistic and analytic ratings (Barkaoui, 2008; Charney, 1984; Cooper, 1984; Cumming, 1990; Gilfert & Harada, 1992; Hamp-Lyons, 1995; Han, 2013; Huang & Han, 2013; Johns, 1991; Lee, 2004; Nakamura, 2004; Razaei & Lovorn; 2010; Rinnert & Kobayashi, 2001; Russikoff, 1995; Shohamy, Gordon, & Kramer, 1992; Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999). Russikoff (1995) strongly claims "holistic assessment may collapse criteria by the one score (and possibly under the weight of one criterion)" (p. 5). Russikoff (1995) reported that when raters previously rated ESL papers holistically, other areas such as "language use" absorbed their attentions; however, after raters completed analytic ratings of the same ESL papers, they were "surprised to see how strong the content and organization of these ESL papers were" (p. 5). These findings indicate that analytic scoring method can be more appropriate for the scoring of ESL writings than holistic scoring (Huang & Han, 2013).

Second, holistic scoring is used in most ESL writing assessment practices although it has received much criticism for its imprecision (Hamp-Lyons, 1995; Russikoff, 1995, Weigle, 2002). Homburg (1984) argues that holistic scoring is appropriate for rating ESL writings, because "with the training to familiarize readers with the types of features present in ESL compositions", holistic scoring "can be considered to be adequately reliable and valid" (p.103). Further, related to the reliability and validity, holistic scoring is still reliable if a detailed rater training is applied and rating session is administrated faithfully and raters strictly adhere to rating criteria (Perkins, 1983; White, 1994); and it also "has the highest construct validity when overall attained writing proficiency is the construct assessed" (Perkins, 1983, p. 652).

Several empirical studies have investigated how the rating methods lead to the variability and reliability of ESL/EFL writing ratings (Barkaoui, 2007, 2008, 2010a, 2010b; Knoch, 2009; Lee et al., 2009). Barkaoui (2007) investigated the effects of holistic and analytic rating scales on EFL essay scores, rating process, and raters' perceptions. Using both scoring scales four EFL writing teachers rated 32 compositions. In qualitative data analysis, think-aloud protocols were collected in the rating of two sets of four papers while *G*-theory approach was used in quantitative data analysis. It was found that there was unexpected higher inter-rater reliability in holistic rating; contrary to what it had been assumed would be found in analytic rating. Yet, the rating processes were alike in both rating procedures.

In the following year, Barkaoui (2008) examined the effects of scoring methods (holistic and multiple trait scales) and different levels of rater expertise on the rating of ESL essays. To illuminate the factors contributing to variability in L2 writing scores, the researcher used quantitative and qualitative approaches in data collection. Each of 31 novice (28 females and 3 males) and 29 (22 females and 7 males) experienced raters rated 24 ESL essays. University level students with varying levels of proficiency in English wrote the 24 ESL essays. Essay scores were analyzed through multi-faceted Rasch measurement and multilevel modeling to observe the interactions between the scores obtained from holistic and analytic ratings. Qualitative data obtained from interviews and think-aloud protocols to examine decision-making behaviors. The results indicated that though both scoring methods measured the same construct, the multiple-trait scoring procedure has able to distinguish more finely among the students' writing abilities. Interestingly, holistic scoring produced higher inter-rater reliability while analytic scoring led to higher rater-self consistency, especially for novice essay raters. While more judgment and self-monitoring strategies were employed in multiple-trait scoring, more interpretation strategies and language focus were observed in the holistic rating process. Furthermore, raters could attend to all rating criteria in the multiple-trait scoring rubric. Intra-rater and inter-rater variability were greater in novice raters' ratings. On the other hand, novice raters

referred more frequently to the rating scale and attended more to the local aspects of writing, besides they spent more time on interpreting and/or editing text; whereas experienced raters referred more to other criteria and spent more time on reading essays to be more self-consistent.

A holistic scoring scale usually has fewer specific descriptors than an analytic scoring scale (Weigle, 2002). More recently, Knoch (2009) compared a scale with fewer descriptors and a scale with more detailed descriptors for writing in an English-for-academic-purpose (EAP) context to find out which scale would result in more valid and reliable ratings. Ten experienced raters rated 100 papers, using two scales. A multi-faceted Rasch measurement analysis was conducted in the quantitative data analysis to compare rater behavior and interviews and questionnaires were used to elicit raters' perceptions about the efficacy of the two rating scales. The results reveal that rater reliability was significantly higher in the scale with a detailed level of descriptors than is the scale with fewer specific descriptors.

Most recently, Barkaoui (2010b) further examined the relationship between the rating scales, rater experience, through think-aloud protocols. Inexpert (n=11) and experienced (=14) raters evaluated 12 ESL essays analytically and holistically. The results revealed that rubric types had more effect on the scoring processes than rater experience. The variations in the rubrics affected inexpert raters more than experienced raters.

In another study, Barkaoui (2010a) used a mixed-methods approach to investigate the variability in the ESL essay holistic scores and evaluation criteria of 32 experienced and 29 novice raters. In this study, holistic and analytic scorings were given by inexpert and experienced raters, with a written explanation for their holistic scores. The quantitative and qualitative data analyses aimed to examine the criteria used while giving holistic scores. The results indicated that the communicative quality of essays was more important than other aspects of writing for both novice and experienced raters; on the other hand, novice raters were more lenient than experienced raters in giving

more importance to argumentation, while the experienced raters were more severe regarding grammatical accurateness.

To sum up, the results of the studies is inconclusive in terms of how the scoring methods affect the variability and reliability. In fact, professionals question the appropriate and effective methodology of scoring ESL/EFL students' compositions for instructional, administrative, and research purposes. This study was designed and conducted to find a solution to this important dilemma. Specifically, using *G*-theory and rater interviews, this study examined the impact of scoring methods (i.e., holistic vs. analytic scoring) on the scoring variability and reliability of an EFL institutional writing assessment at a Turkish university.

The *G*-theory Approach

The *G*-theory (Cronbach et al., 1972) approach measures the dependability of behavioral measurements (Webb & Shavelson, 2005). *G*-theory is an extension of classical test theory (*CTT*), which provides a single estimate of error at a time; however, *G*-theory is the expansions of the *CTT* to separately estimate the several sources of error affecting test scores (Shavelson & Webb, 1991). It can be used to examine the relative contribution of multiple sources of error as well as their interactions on the generalizability of the assessment results (Shavelson & Webb, 1991). The conceptual framework of *G*-theory differs from CTT in several respects. First, *G*-theory examines the multiple sources of variability simultaneously. Second, *G*-theory estimates the magnitude of each source of variance. Third, *G*-theory calculates two different reliability coefficients (*Phi*-coefficient and *G*-coefficient).

As described by Shavelson and Webb (1991), *G*-theory extends the analysis of variance (ANOVA) approach of *CTT* to reliability. The estimation of the components of variance is not unique to *G*-theory. ANOVA can provide an estimation of the components of variance and then a calculation of reliability. It is possible to estimate the magnitude of important independent variables through ANOVA. In a random ANOVA analysis, only a single source of error can be considered at a

time. For example, when the error source is *occasion,* the score for each individual on each occasion would be summed over items; likewise, *items* could be considered as a source of error (Shavelson & Webb, 1991).

Research Questions

Four research questions were asked in this study: a) Are there significant differences between the holistic and analytic scores of the same EFL paper? b) What are the sources of score variation contributing relatively more to the score variability of the holistic scores in contrast to the analytic scores assigned to the EFL papers? c) Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations) of the holistic scores differ from the analytic scores assigned to the EFL papers? d) What is the impact of scoring methods (holistic vs. analytic scoring) on raters' decision making during the rating processes?

Methodology

Data Collection Procedures

The writing samples were taken from the English Language and Literature Department of a state university in Turkey. Data were collected at several steps. Prior to getting the data, permissions were received from the Dean's Office of the Faculty of Letter and Sciences of the University; further related permissions from English Language and Literature (ELL) Department were obtained. Furthermore, argumentative pen-paper-based essays written by EFL students who took the institutional undergraduate English examinations were selected. In addition, all teachers at the ELL department were first explained the context of the study and later requested to participate in the study as volunteer participating raters. The data collection included the actual rating of the EFL essays by the ten volunteer raters, which took place in the fall semester of the 2011-2012 academic year.

The Selection of Writing Samples

The selection of the writing samples was undertaken as follows. Initially, all English instructors within the ELL department at the university were first invited to partake in the study. Then four instructors were randomly appointed to select students' English writing samples from the institutional undergraduate English examinations for data analysis. The writing component of the examination required undergraduate students to write an argumentative essay on one prompt in 45 minutes.

Each instructor selected nine argumentative essays written by nine undergraduate Turkish-speaking students in his or her class. These nine papers were evaluated by the instructor as representing three different levels of quality (high, medium, and low) in order to maximize the differences among papers. Totally, 36 papers were selected for this study. In this study, ten raters scored these 36 papers first holistically and then analytically, with a three-week time interval. This resulted in 36 papers written by 36 persons (*p*), each paper or person receiving twenty different scores (i.e., ten holistic and ten analytic scores) from ten raters (*r*) through holistic and analytic scoring methods (*m*).

The Selection of Raters

The ten participating raters were five males and five females. They were volunteer lecturers, research assistants, and university professors with a various teaching background. They had minimum one year of experience in teaching and assessment. Their mother tongue was Turkish and they were all proficient EFL speakers. Their ages ranged from 20 to 50. It was interesting to note that eight out of the ten raters frequently used the holistic scoring method in marking EFL essays.

The Rating Scales

The instruments used in the study were the department holistic and analytic scales that were modified by the researchers according to the literature on rubric development (Brown, 2004a; East, 2009;

Russikoff, 1995; Weigle, 2002), the Turkish EFL students' writing samples, rater and faculty input, and assessment objectives. The holistic rubric was a 10-point holistic scale that included the following writing performance criteria: a) grammar, b) content, c) organization, d) style and quality of expression, and e) mechanics.

The holistic rubric was incorporated to the descriptors in the analytic rubric. It is important to note that the "mathematical assignment" of the 10-point analytic scale did not give equal weight to each category but they were weighted; that is, different point values were given in the five categories. Table 1 shows the five weighted categories and the point values in the 10-point analytic scale used in the study.

Table 1: The Score Weights of Five Categories in the 10-point Analytic Scale

Category	Weight Percentage
Grammar	30 %
Content	20 %
Organization	20 %
Style and quality of expression	15 %
Mechanics	15 %

The Rater Training and Rating Procedures

Training raters are essential to obtain reliable results. Therefore, raters are calibrated so that they apply the same standards to their scoring through training (Lenel, 1990). All raters in this study received a thorough training before scoring the writing samples. One of the researchers participated as a rater trainer in the study after receiving rater training from external experts. A traditional classroom model was applied to train raters based on the rating method used. The two characteristics of this model include a) it was conducted in a group setting, trainer as a teacher, raters as students, and b) the format was pen-paper based; therefore it was cost effective (Johnson et al., 2009). As Johnson et al. (2009) emphasized that the procedures and materials used to familiarize raters with the process, the criteria used in the

assessment, and the task(s) should be explained while training raters to rate performance. For this study, a rater-training plan was adapted from Barkaoui (2008) and Johnson et al. (2009).

In each training session (i.e., the rater training for holistic and analytic methods), the purpose and the context of the study were described by the first author of this study to the raters. Further, each rating scale was reviewed and explained with a focus on the descriptors and the writing task. Then, the researcher allotted several minutes for discussion period. In this period, each rating scale was discussed with the raters in terms of the expectations and writing tasks in the study. Following that, the scoring of essays of different quality (i.e., good, average, poor) by using both scales was modeled followed by the raters' practice rating with a sample of nine essays with different qualities. Finally, the raters discussed their ratings, and negotiated and solved their disagreements if there were any.

In this study, the training sessions were conducted in the same manner for holistic and analytic ratings using the same techniques and the same practice papers. All the participants attended two training sessions and each lasted approximately 25 minutes. There was a three-week time interval between the two training sessions.

In the first holistic scale training session, each rater was trained to interpret the scoring dimensions in the holistic rubric. After that training session, a small-scale pilot study which included raters' ratings and discussions on a randomly selected sample of nine papers was conducted to make sure that the raters understood the holistic rubrics. Then, the raters received a package containing the holistic scoring rubric and the 36 EFL essays. Each rater read each composition quickly and then individually judged against the 10-point holistic scoring rubric. The aim of holistic rating was to rate the overall proficiency level reflected in the sample of 36 EFL student compositions.

The second stage started three weeks after the holistic scoring. Each of the same raters was trained for another 25 minutes to use the analytic rubric. Immediately after this training session, the raters scored a different sample of nine papers by using the analytic rubric.

They then discussed their evaluations to reduce score variations. For this session, the raters were given the analytic scoring rubric and the same 36 EFL essays in a parcel. At the end of the second training session, each rater scored the same 36 argumentative essays analytically. In each rating session, they scored the papers individually to prevent discussion amongst the raters.

It is important to note that the raters were not told that they had scored the same essays holistically before. Further, a three-week time interval was assumed to be sufficient for raters to forget about the scores they had assigned to these essays. Practice papers were not included in the actual study to prevent familiarity.

Finally, face-to-face interviews were conducted with a random sub-sample of four raters (*A*, *B*, *C*, and *D*) from the ten participating raters right after each rating session. Interview questions were prepared and directed to the raters after their holistic and analytic scoring sessions, respectively. The interview questions focused on the impact of scoring methods (i.e., holistic vs. analytic scoring) on their scoring decision making during the rating processes. Specifically, the interviews investigated the raters' perceptions of each scoring method as well as the factors most impacting their holistic versus analytic scoring decisions.

The Data Analysis Methods

First, descriptive statistical analyses and *t*-tests for the holistic and analytic writing scores given by the ten raters were conducted. The purpose of conducting these analyses was to investigate whether any significant mean score differences existed between the holistic and analytic scores assigned by the ten raters in the study.

Within *the G*-theory framework, further data analyses were done in the following three steps: 1) person-by-method-by-rater random effects *G*-study; 2) person-by-rater random effects *G*-studies for the holistic and analytic scoring, respectively, and 3) calculation of dependability coefficients.

Person-by-method-by-rater Random Effects G-studies

In this study, ten raters scored all 36 papers first holistically and then analytically. This resulted in 36 persons (*p*) and 720 scores, each person receiving twenty different scores (i.e., ten holistic and ten analytic scores) from ten raters (*r*) through holistic and analytic scoring method (*m*). Therefore, this constitutes a fully crossed *person-by-method-by-rater* (*p x m x r*) *G*-study design. This *G*-study analysis was aimed to obtain variance component estimates for the seven independent sources of variation: person (*p*), rater (*r*), method (*m*), person-by-rater (*p x r*), person-by-method (*p x m*), method-by-rater (*m x r*), and person-by-rater-by-method (*p x r x m*). (cf.Han & Ege, 2013)

Person-by-rater Random Effects G-studies

Additionally, two separate paper-by-rater (*p x r*) random effects *G*-studies were conducted for the holistic and analytic writing scores, respectively. These *G*-studies aimed to compare the holistic and analytic scores in terms of score variability and reliability. With the implementation of these *G*-studies, the three independent sources of variation, namely, person (*p*), rater (*r*), and person-by-rater (*p x r*) for each scoring method were obtained. By the use of the obtained variance components, dependability coefficients for each scoring method were then calculated for examining the reliability (cf. Han, 2013; Huang, 2012).

Finally, a coding and classifying approach (Gay, Mills, & Airasian, 2009) was used for rater interview data analysis. The raters' responses pertinent to the last research question were categorized and analyzed according to the recurring themes.

Computer Programs

Descriptive and inferential statistical analyses were performed with SPSS. The G-study analyses were performed with the computer program GENOVA (Crick & Brennan, 1983).

Results

Descriptive Statistical Results

Table 2 provides the descriptive statistics for the holistic and analytic scoring data used in the analysis.

Table 2: Descriptive Statistics for Both Scoring Methods

Paper	Holistic Scoring		Analytic Scoring		Mean Difference (A-H)*
	Mean	SD	Mean	SD	
1	5.50	1.41	6.54	1.54	1.04
2	6.60	1.90	6.27	1.16	-0.33
3	5.70	0.92	6.03	1.41	0.33
4	5.65	1.18	6.33	1.38	0.68
5	5.85	1.78	6.20	2.01	0.35
6	6.80	1.27	6.33	1.60	-0.47
7	6.15	1.81	5.88	1.70	-0.27
8	6.90	1.43	5.75	1.43	-1.15
9	3.55	1.46	3.80	1.37	0.25
10	3.50	1.13	4.12	0.90	0.62
11	6.90	2.02	7.34	1.03	0.44
12	7.05	1.21	6.48	1.46	-0.57
13	5.95	2.29	6.50	1.63	0.55
14	5.45	1.17	5.71	1.23	0.26
15	4.20	1.38	3.48	1.08	-0.72
16	4.45	1.71	4.53	1.06	0.08
17	5.55	1.54	4.47	1.11	-1.08
18	5.45	1.67	4.99	1.36	-0.46
19	4.90	2.05	4.19	1.50	-0.71
20	2.15	0.85	2.91	1.35	0.76
21	3.80	1.46	3.70	1.58	-0.1
22	3.10	1.20	3.73	1.79	0.63
23	3.75	1.51	4.31	1.50	0.56
24	4.45	1.44	4.93	1.02	0.48
25	4.70	1.77	4.35	1.00	-0.35
26	5.30	1.34	5.05	1.43	-0.25
27	3.50	0.97	3.62	1.78	0.12
28	3.30	1.30	3.75	1.28	0.45
29	2.55	0.93	2.94	0.70	0.39
30	5.65	1.33	6.01	1.24	0.36
31	6.75	1.11	6.20	1.80	-0.55
32	5.80	1.62	5.95	1.67	0.15
33	6.10	1.41	5.84	1.28	-0.26
34	4.90	1.66	4.49	1.03	-0.41
35	6.25	1.51	6.78	1.66	0.53
36	6.65	1.33	6.70	1.46	0.05

Note: N (rater) = 10; *Mean Difference (A-H) = Mean Difference (Analytic Score – Holistic Score)

When the holistic and analytic scores were compared, the results show that 21 out of 36 papers received higher scores for analytic scoring than for holistic scoring; 15 out of 36 papers received higher scores for holistic scoring than for analytic scoring; further the mean score difference for only one paper (i.e., paper #1) was greater than one score point, with higher analytic scores than holistic scores for that paper. Interestingly, the mean score difference for two papers (papers #8 and #17) was greater than one score point, with higher holistic scores than analytic scores for the two papers. Further, for holistic scoring 32 out of 36 papers had a standard deviation of over one score point; similarly, for analytic scoring 34 out of 36 papers had a standard deviation of over one score point, indicating that there was great rater variation of both holistic and analytic scoring of these EFL papers.

The descriptive statistical results suggest that holistic and analytic scoring methods yielded similar results. In other words, the scoring methods did not have much impact on the EFL writing scores.

Inferential Statistical Results

Paired sample *t*-tests for both scoring methods was conducted to examine the significant mean score difference between the holistic and analytic scores assigned by the ten raters.

As shown in Table 3, there was a significant difference between the holistic and analytic scores for only paper #8 ($p < .05$). The holistic score this paper received was significantly higher than the analytic score it received. For all other papers, there was no significant mean score difference between holistic and analytic marking.

Similarly, the inferential statistical results and the descriptive statistical results were complementary, this means that as holistic and analytic scoring methods yielded similar results, the scoring methods did not have much impact on the marking of the EFL essays.

Table 3: Paired Samples t-Tests Results

Pair	DF	*t*	Sig.
Pair 1: HL-AN (Paper #1)	9	-16.66	.13
Pair 2: HL-AN (Paper #2)	9	0.53	.61
Pair 3: HL-AN (Paper #3)	9	-0.83	.43
Pair 4: HL-AN (Paper #4)	9	-15.71	.15
Pair 5: HL-AN (Paper #5)	9	-0.36	.73
Pair 6: HL-AN (Paper #6)	9	0.69	.51
Pair 7: HL-AN (Paper #7)	9	0.56	.59
Pair 8: HL-AN (Paper #8)	9	2.62	.028*
Pair 9: HL-AN (Paper #9)	9	-0.51	.62
Pair 10: HL-AN (Paper #10)	9	-1.64	.14
Pair 11: HL-AN (Paper #11)	9	-0.68	.51
Pair 12: HL-AN (Paper #12)	9	13.55	.21
Pair 13: HL-AN (Paper #13)	9	-0.79	.45
Pair 14: HL-AN (Paper #14)	9	-0.54	.60
Pair 15: HL-AN (Paper #15)	9	13.80	.20
Pair 16: HL-AN (Paper #16)	9	-0.15	.89
Pair 17: HL-AN (Paper #17)	9	21.28	.06
Pair 18: HL-AN (Paper #18)	9	13.77	.20
Pair 19: HL-AN (Paper #19)	9	13.14	.22
Pair 20: HL-AN (Paper #20)	9	-14.29	.19
Pair 21: HL-AN (Paper #21)	9	0.19	.86
Pair 22: HL-AN (Paper #22)	9	-12.95	.23
Pair 23: HL-AN (Paper #23)	9	-11.95	.26
Pair 24: HL-AN (Paper #24)	9	-14.76	.17
Pair 25: HL-AN (Paper #25)	9	0.95	.37
Pair 26: HL-AN (Paper #26)	9	0.58	.57
Pair 27: HL-AN (Paper #27)	9	-0.23	.83
Pair 28: HL-AN (Paper #28)	9	-10.82	.31
Pair 29: HL-AN (Paper #29)	9	-11.24	.29
Pair 30: HL-AN (Paper #30)	9	-0.62	.55
Pair 31: HL-AN (Paper #31)	9	0.78	.45
Pair 32: HL-AN (Paper #32)	9	-0.19	.86
Pair 33: HL-AN (Paper #33)	9	0.54	.60
Pair 34: HL-AN (Paper #34)	9	0.74	.48
Pair 35: HL-AN (Paper #35)	9	-0.95	.37
Pair 36: HL-AN (Paper #36)	9	-0.10	.92

Note: *indicates significant difference at the .05 level; HL: holistic score; AN: analytic score.

G-study Results of Person-by-method-by-rater Random Effects

The person-by-method-by-rater (*p x m x r*) random effects *G*-study is shown on Table 4.

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

Source of Variability	DF	σ^2	%
p	35	1.4599	40.80
m	1	0.0	0.00
r	9	0.1793	5.01
pm	35	0.0300	0.84
pr	315	0.4458	12.46
mr	9	0.3180	8.89
pmr	315	1.1448	32.00
Total	719	3.5778	100

As shown in Table 4, the following seven variance components produced respectively:

1) Person (*p*), the object of measurement yielded 40.8% of the total variance. This result reveals that the 36 EFL papers were substantially different concerning quality
2) The residual yielded 32% of the total variance. This suggests the interaction between raters, scoring methods, persons, and other unexplained systematic and unsystematic sources of error.
3) Person-by-rater (*pr*) yielded 12.46% of the total variance. This result reveals that raters marked all papers very differently
4) Method-by-rater (*mr*) yielded 8.89% of the total variance. This result demonstrates that the inconsistency concerning rating severity or leniency across scoring methods is very high.
5) Rater (*r*) yielded the fifth largest variance component (5.01% of the total variance). This result shows that raters differed from one another concerning leniency of scoring these papers.

6) Person-by-method (*pm*) yielded only 0.84% of the total variance). This result shows that these papers are relatively similar concerning scores across scoring methods.
7) The variance component for scoring method (*m*) did not explain any total score variance. This result suggests that there was not much difference in the writing scores stemming from the scoring method itself.

G-studies Results regarding Person-by-rater Random Effects

Two separate person-by-rater (*p x r*) random effects *G*-studies were performed for the holistic and analytic scores, respectively. These *G*-studies was aimed to compare the holistic and analytic scores concerning score variability and reliability. Table 5. shows the results.

Table 5: Variance Components for Random Effects p x r G-study Designs

Scoring Method	Source of Variability	DF	σ^2	%
Holistic Scoring	*p*	35	1.6290	42.54
	r	9	0.6746	17.62
	pr	315	1.5258	39.84
	Total	359	3.8294	100
Analytic Scoring	*p*	35	1.3507	40.61
	r	9	0.3200	9.62
	pr	315	1.6554	49.77
	Total	359	3.3261	100

Table 4 shows the following three variance components produced respectively for each scoring method:

1) The results for the holistic scoring method show that person (*p*) yielded 42.54% of the total variance which is the largest variance component. This result suggests that the 36 EFL papers were considerably different in terms of quality.
2) The residual yielded 39.84% of the total variance. This second largest variance component comprises the variability because of the interaction between raters and papers, and other unexplained systematic and unsystematic sources of error.

3) Rater *(r)* yielded 17.62% of the total variance. This third largest variance component indicates that raters did differ considerably from one another in terms of leniency of marking these EFL papers.

Again as Table 5 presents:

1) The results for the analytic scoring method show that the residual yielded 49.77% of the total variance which is the largest variance component.
2) Person *(p),* yielded 40.61% of the total variance. This second largest variance component indicates that the 36 EFL papers were considerably different in terms of quality.
3) Rater *(r)* yielded 9.62% of the total variance. This third largest variance component indicates that raters did differ considerably from one another concerning leniency of scoring these EFL papers.

It is important to point out that the variance component due to rater *(r)* was approximately two times bigger for holistic scoring (17.62% of the total variance) than for analytic scoring (9.62% of the total variance). This result indicates that there was considerably more rater inconsistency for holistic scoring than for analytic scoring in terms of scoring leniency.

In summary, the results of the person-by-method-by-rater random effects *G*-studies and the paper-by-rater random effects *G*-studies suggest that the great source of score variation for both holistic and analytic scoring methods stemmed from the differences among EFL students' EFL writing performance as measured by the writing tasks. This result confirms that writing tasks did distinguish among EFL students if either method was used. Further, large residual variance component in both scoring methods indicated that other facets (e.g., quality of EFL essays), which may have attributed to the score variance, were not considered in the design (Brennan, 2001). The fairly large variance component of method-by-rater *(mr)* suggests that rater inconsistency between scoring methods did contribute to the EFL

writing score variance. Further, the variance component due to rater *(r)* was approximately two times bigger for holistic scoring than for analytic scoring. This result indicated that there was considerably more rater inconsistency concerning scoring leniency for holistic scoring than for analytic scoring.

Dependability Coefficients

The dependability coefficients for each scoring method were calculated and presented in Table 6.

Table 6: Summary of Dependability Coefficients

		Dependability Coefficients	
Number of Papers	Number of Raters	Holistic Scoring	Analytic Scoring
36	1	.43	.41
36	2	.60	.58
36	3	.69	.67
36	4	.75	.73
36	5	.79	.77
36	6	.82	.80
36	7	.84	.83
36	8	.86	.85
36	9	.87	.86
36	10	.88	.87
36	11	.89	.88
36	12	.90	.89
36	13	.91	.90
36	14	.91	.91
36	15	.92	.92
36	16	.92	.92
36	17	.93	.92
36	18	.93	.92
36	19	.93	.93
36	20	.94	.93

Table 6 shows the dependability coefficient calculated for the holistic scoring method for the current ten-rater scenario was .88; the dependability coefficient for the analytic scoring method was .87. Additionally, the result indicates if the number of raters was increased

to 20 for the holistic scoring method, the dependability coefficient would result in .94, almost identical to the dependability coefficient of .93 which was obtained for the analytic scoring when the number of raters was 20.

In summary, the dependability coefficients that were obtained for holistic scoring are very similar to those obtained for analytic scoring of the EFL essays. This result indicates that the scoring method does not have a significant impact on the rating reliability of the EFL essays. In other words, holistic scoring could yield reliable and dependable results as analytic scoring.

Rater Interview Results

Face-to-face rater interviews were conducted for the purpose of examining the impact of each scoring method on raters' scoring decision making. The data derived from the follow-up interviews with the four raters were used to answer the last research questions, i.e., what is the impact of scoring methods (i.e., holistic vs. analytic scoring) on their decision making during the rating processes? The data obtained from rater interviews were grouped together and then categorized into the following two themes: a) raters' perceptions of each scoring rubric, and b) the factors most impacting their holistic versus analytic scoring decisions.

Raters' Perceptions of Scoring Methods

All four raters agreed that the holistic scoring method was both useful and beneficial in scoring these EFL essays. The holistic rubric makes the scoring fair and objective. Rater B further commented that

> *"… I used to evaluate the student papers according to my intuition or experience [subjective criteria] … [but now] in here when I used the holistic rubric to score essays, I had to be objective; which was a good thing. I am a fair scorer now."*

Further, all four raters agreed that the holistic scoring rubric had several strengths. Two male raters, A and B, reported that the holistic scoring rubric helped the raters assign fair and objective scores to essays. Two female raters, C and D, believed that the holistic scoring rubric facilitated their scoring process.

In contrast, all raters mentioned that the analytic scoring method could help get a more reliable score because it considers all aspects of an essay. However, they all reported that using the analytic scoring method was very time-consuming. They had to read the rubric frequently while marking each essay. Raters C commented that *"It's challenging if we have to consider "time". I mean that it is time-consuming. Scoring becomes very difficult when there are over fifty papers."* Rater D further commented that: *"It's challenging if we have to consider "time". I mean that it is time-consuming. Scoring becomes very difficult when there are over fifty papers."*

To sum up, as reported by the raters, each scoring method has both strengths and weaknesses. The holistic scoring rubric was liked by all four raters. It can help them not only assign fair and objective scores to essays but also facilitate their scoring process.

Factors Most Impacting Scoring Decision Making

For holistic scoring, Raters A, B, and C agreed that the content of an essay was the most important factor that affected their scoring decision making of the essay. For example, Rater A made the following comment:

> *"… the content of the paper are [is] more important, but we should also pay attention to its grammar use … grammar is in the second place for me. Content is the most important factor that impacted my scoring of each paper."*

However, Rater D indicated that the grammar of an essay affected her scoring decision most. She commented that "*The organization is important but grammar is the most important …."*

For analytic scoring, due to the nature of the analytic scoring rubric, all raters agreed that they considered all aspects of an essay in

making their scoring decisions. In other words, a single aspect of the writing (e.g., grammar, content, or organization) did not influence their ratings with the analytic scoring. However, Rater D mentioned that she would consider the language and content of an essay most in using the analytic rubric to mark the essay. The following comment was made by Rater D:

> *"The things that most impacted me while scoring the papers [analytically] were the language and content. These two are [most] important because they give the paper a meaning – the thing that the student actually tries to put forward in his [or her] paper – but these two [i.e., language and content] cannot do without organization, capitalization and other aspects."*

Discussion and Conclusions

The first research question was asked to obtain information if there would be any difference between the holistic and analytic scores of the same EFL paper. Descriptive statistics and inferential statistics were conducted to answer this research question. The descriptive statistical results showed that 21 out of 36 papers received higher scores for analytic scoring than for holistic scoring. Previous literature indicates that scoring each dimension (grammar, content, organization, etc.) separately in analytic scoring may lead the rater to give higher scores than in holistic scoring (Alharby, 2006; Barkaoui, 2008; Goulden, 1994). This result appears to be consistent with the previous studies.

However, for both holistic and analytic scoring, nearly all papers had a standard deviation of over one score point. This result indicated that there was great rater variation for both holistic and analytic scoring of these EFL papers. Rater variation in holistic scoring "may not be explained in terms of the criteria in the rating scale only" (Barkaoui, 2008; p.18) because raters may use their personal judgment and give importance to a single different criterion out of the criteria listed in the holistic scale (Goulden, 1994). Further research is needed to examine the great rater variance in analytic scoring.

Further, the paired sample *t*-tests for the holistic and analytic writing scores were conducted to investigate closely whether there was a significant mean score difference between the holistic and analytic scores. The results showed that except for a single paper (i.e., paper #8, $p < .05$) there was no significant mean score difference between holistic and analytic marking for all other papers. In other words, the inferential statistical results indicated that holistic and analytic scoring methods yielded similar results; and therefore, the scoring methods did not have much impact on the scoring of these EFL essays. This result has been confirmed by several research studies (Gilfert & Harada, 1992; Lee, 2004; Nakamura, 2004; Song & Caruso, 1996) although it is contradictory to the findings as reported by Russikoff (1995), who claimed that analytic scoring generally considers all aspects of a piece of writing and result in less rater variation and more reliable writing scores than holistic scoring.

One reason for the similar results across holistic and analytic scoring may be because of the detailed rater training, which might have alleviated the score differences and increased the scoring reliability. Interestingly, in a study on the reliability and validity of scoring rubrics Razaei and Lovorn (2010) found that using rubrics may not improve the reliability or validity of assessment without rater training but if raters are well trained on how to design and employ them effectively, more reliable results can be obtained. Therefore, if an intensive rater training program for scoring writing performance is applied, rater consistency may be improved because training provides the rater with a clear conception of what a piece of quality writing is (Shohamy, et al., 1992). The literature reported that rater training can effect on applying the scoring criteria on the rubric reliably and, thus, it increases the reliability of the interpreting and scoring dimensions of the scoring scale (Stuhlmann, et al., 1999).

The second research question intended to examine the differences in score variations between holistic and analytic scores assigned to each essay. The data were further analyzed to obtain information for comparison between holistic and analytic scores in terms of score variability and reliability, using *the G*-theory framework.

The results of *G*-studies indicated that the great source of score variation for both holistic and analytic scoring methods stemmed from differences among the students' EFL writing skills as measured by the writing tasks. The desired variance associated with the object of measurement (i.e., persons) for both methods was similar. These results confirm the writing tasks did distinguish among EFL students if either method was used. However, the greatest source of score variation for both scoring methods was due to the residual variance component, not differences among the students' EFL writing performance as measured by the writing tasks. The large residual variance component indicated that other facets (e.g., quality of EFL essays), which may have attributed to the score variance, were not considered in the design (Brennan, 2001).

The third research question examined the differences in the reliability between holistic and analytic scores assigned to EFL papers. The dependability coefficients for each scoring method were calculated. The results showed that the dependability coefficient obtained for holistic scoring was very similar to that obtained for analytic scoring of the EFL essays. Even increasing the number of raters to 20 for each method, the scenario would be identical. This result indicates that the scoring method does not have a significant impact on the rating of the EFL essays. So, holistic scoring was able to produce reliable and dependable results as analytic scoring.

The above results may be explained by the detailed rater training applied to the raters in the study. Training raters are essential in order to score a piece of writing consistently (Shohamy et al., 1992; Weigle, 1994). Several studies have investigated the effect of rater training on essay scores (Stuhlmann, et al., 1999; Weigle, 1994, 1998). The results of these studies indicated that the training of raters can be important in terms of reliability and, hence, the validity of the ratings.

The last research question asked about the impact of scoring methods (i.e., holistic vs. analytic scoring) on four raters' decision making during the rating processes. As reported by the raters, although each scoring method has both strengths and weaknesses, they all liked the holistic scoring method. It could help them not only

assign fair and objective scores to essays but also facilitate their scoring process. Further, most raters agreed that the content of an essay was the most important factor that most affected their holistic scoring decision making of an essay. In contrast, all aspects of an essay (e.g., grammar, content, or organization) jointly affected their analytic scoring decision making of an essay.

There are several limitations that need to be acknowledged and addressed regarding the present study. First, this study included only one authentic argumentative essay by EFL students. Research has shown that different writing tasks (narrative essays, argumentative essays, etc.) affect the scoring variability and reliability of ESL/EFL essays (Huang, 2008, 2012).

Second, the order of scoring methods and the organization of rating scales might have impacted the results of the study. For example, the results might have been different if the raters of this study had scored the EFL papers first analytically and then holistically or the levels in the scales were implemented differently. As Barkaoui (2008) stated, the organization of rating scales may impact the scores assigned to writing samples. In this study, both holistic and analytic scales list grammar and content as the top two criteria for scoring. The inexperienced raters could have tended to attend more to grammar and content and judge the overall quality of essays with both scales depending on one of these single criteria while scoring these essays (Rinnert & Kobayashi, 2001).

The third limitation may be related to the rating procedure and context. The ratings were done at raters' homes or in their offices with a three-week time interval between the holistic and analytic scoring sessions. The flexible scoring procedure might have impacted the scores the raters assigned to the EFL essays. Further, although a three-week time interval between the two scoring sessions was thought to be sufficient, it may not be long enough for the raters to forget about their scores assigned to these papers because raters may recognize the essays when they marked the second time (Barkaoui, 2008).

Finally, the use of rater interviews instead of rater think-aloud protocols might have limited the qualitative results of this study and

the interpretation of these results. Although rater follow-up interviews are a viable alternative to think-aloud protocols, due to the data collection time (during vs. after rating), there was the probability of not remembering all aspects of the rating process. The changing behaviors of the raters in the rating process could affect the data quality and result in a discrepancy between what was reported in the score analysis and what was reported in the interviews (Barkaoui, 2008). On the other hand, think-aloud protocol analysis is very effective in providing the "richest evidence" about how raters behave while scoring ESL/EFL compositions; therefore, the research on the scoring process can focus on both rating scale validity issues and fairness issues (Connor-Linton, 1995; Weigle, 1994).

The following four conclusions were reached based on the limitations: First, the effects of both the rater training and detailed scales might have masked the differences in score reliability between the holistic and analytic scoring. In this study, raters were not only oriented to use the scoring rubrics skillfully but also received a detailed rater training in using both rating scales consistently. Further, most of the participating raters had experience less than two years in scoring EFL writing; an effective rater training could increase sensitiveness and yield this higher consistency between ratings (Weigle, 1994). Findings from the study suggested that when a detailed rater training is applied, the holistic rating could be used to ensure as fair and consistent ratings as the analytic rating.

Second, although there was little difference in score reliability between the holistic and analytic scoring, there was considerably more rater inconsistency in terms of scoring leniency for holistic scoring than for analytic in the study. The variance component due to rater was approximately two times bigger for holistic scoring than for analytic scoring. This reveals that holistic evaluations may be unduly influenced by superficial features of the writing samples although rating sessions are supervised carefully; thus more serious attention to the validity of the scores should be given in holistic scoring (Charney, 1984).

Third, there is a great unexplained variability. The greatest source of score variation for both scoring methods was due to the residual variance component. The residual comprises the variability stemming from the interaction between raters and persons, and other unexplained sources of error whether systematic and unsystematic. Further, large residual effects may be an indication of 'hidden facets' that may have attributed to the score variance (Brennan, 2001). "The variance of the hidden facets is included in the residual variance, thus leading to a larger residual than when the facet is explicitly considered" (Huang, 2008, p.215).

Finally, the study showed that holistic scoring method could achieve comparable reliability as analytic scoring. Using holistic scoring method, a large number of students' writing performance could be assessed for the purpose to place them into different levels of writing courses in less time and cost (Weigle, 2002). Further, Cumming (1990) "advocated holistic scoring because comparable results can be achieved more quickly than with analytical approaches" (quoted in Johns, 1991, p. 380). Therefore, for practicality and cost-effectiveness (e.g., less time taking, less labor, etc.), holistic assessment can also be used reliably in placement writing course exams, diagnostic exams and exit exams both in large-scale and small-scale (e.g., classroom assessment) context; as this study puts forwards, if the raters are appropriately trained in using rubrics specifically developed for the specific EFL students to be assessed depending on the former experiences with the students' writing samples.

This study was intended to investigate the impact of scoring methods (holistic vs. analytic scoring) on the variability and reliability of EFL writing assessments. The results have implications for EFL writing assessment practices.

The following three recommendations are recommended based on the findings of this study: First, EFL writing assessment professionals in the institutional assessment context as well as EFL writing course teachers in the classroom assessment context should develop a detailed, clear, and easy-to-follow scoring guide for either holistic or analytic scoring. This scoring guide should be developed

with the contribution of all the writing course teachers involved. The development of a clear and easy-to-follow scoring guide and the appropriate use of the guide in terms of applying the criteria when marking EFL compositions, the inconsistencies between the holistic and analytic scorings could be minimized (Huang, 2010).

Second, EFL writing assessment professionals in the institutional assessment context should provide sufficient and systematic rater training. The findings of this study together with findings from previous studies do support strong recommendations on the essentiality of applying rater training to obtain more consistent and reliable score irrespective the scoring method and rater experience (Barkaoui, 2008; Weigle, 1998; 2002). This study also has implications for assessing EFL essays in terms of using holistic or analytic scoring methods in institutional high-stakes writing assessments and applying rater training. Few studies have been conducted to investigate the impact of the scoring methods on the variability and reliability in EFL writing assessment context. Findings from the study suggested that when a detailed rater training is applied, the holistic rating could be adopted to ensure fair and consistent ratings as the analytic rating.

Finally, it is suggested that a well-developed holistic scoring rubric be used in the institutional assessment context, in which the assessment is for placement and diagnosis purposes. Using holistic scoring will make the assessment not only less time-consuming and cost-effective than using the analytic scoring method but also as unbiased as analytic scoring (Weigle, 2002). Holistic scoring is quicker, easier, cost effective and, moreover, the holistic rating is preferable in norm-referenced testing situations where a student's writing performance is compared to other students' performances and therefore it carries "less diagnostic and pedagogical value" (Cooper, 1984, pp. 34-35).

The current study has provided at least three different avenues for future research in the area of EFL writing assessment. First, future research should consider the quality of essays, rating category, and the rater educational background and experience in marking and teaching EFL writing as factors in the analysis because these factors might

impact the scoring of EFL papers (Alharby, 2006). Moreover, more complex studies of the effects of the numbers of raters, task types, rating categories, occasions, and others would be very beneficial and timely, using G-*theory* (Brown, 2004b).

Second, it is suggested that future research use think-aloud protocols instead of raters' follow-up interviews in order to examine raters' scoring decision-making processes. This is because the think-aloud protocol analysis can provide the "richest evidence" about what raters think and do while rating EFL essays (Connor-Linton, 1995; Cumming, 1990; Weigle, 1994).

Finally, future research could include both norm-referenced and criterion-referenced testing programs, use both *G*-theory and other sophisticated models such as item response theory approaches (e.g., multi-facet Rasch measurement), structural equation modeling to examine the factors that affect the assessment of EFL students' English writing performance (Huang, 2008), and use Rasch many-faceted measurement to investigate the performance of holistic and analytic scales and to determine whether the scales separate students' ESL/EFL writing performance. Such models can help us further understand EFL writing assessment issues thoroughly (Huang, 2011).

The authors' note*: This study is a part of the doctoral dissertation by Han (2013). The second author served as the supervisor of this dissertation.*

The Authors

Turgay Han is an assistant professor at the Department of English Language and Literature, Faculty of Letters of Kafkas University (2006 - 2016) and Ordu University (2016 - present). His areas of research interest center on EFL measurement and assessment issues, and his areas of scholarship include assessing language skills, and using G-theory to examine score variability and reliability of EFL writing assessments.

Jinyan Huang (Ph.D.) is a Professor of Leadership and Policy at Niagara University. He earned his Ph.D. (2007) in measurement, assessment and quantitative research methods from Queen's

University at Kingston in Canada. As part of his Ph.D. program, he studied at the Centre for Research in Applied Measurement and Evaluation (CRAME) (2004) at the University of Alberta in Canada. Dr. Huang's areas of research center on large-scale assessment, leadership, and policy issues. Specifically, he is interested in the following four issues: a) factors or level of factors that affect students' large-scale standardized test scores; b) assessment issues (reliability, validity, and fairness) in schools and universities; c) leadership traits and leader effectiveness in organizations.

References

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic, using two measurement models, the generalizability theory and the many-facet Rash measurement, within the context of measurement of performance assessment.* (Unpublished Doctoral Dissertation). The Pennsylvania State University, University Park, PA, U.S.A.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.* (Unpublished Doctoral Dissertation). University of Toronto, Toronto, Canada.

Barkaoui, K. (2010a). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Brennan, R. L. (2001). *Generalizability theory.* Iowa: ACT Publications.

Brown, H. D. (2004a). Language assessment: Principles and classroom practices. White Plains, NY: Pearson Education.

Brown, J. D. (2004b). Performance assessment: Existing literature and directions for research. *Second Language Studies, 22*(2), 91-139.

Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207-241.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research* (GRE Board Research Report, GREB No: 82-15R). Retrieved June 26, 2011, from http://www.ets.org/research/policy_research_reports/rr-84-12

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENeralized Analysis of VAriance system* (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.

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

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R. & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171-176.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing writing, 14*(2), 88-115.

Ebel, R., & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliff, NJ: Prentice-Hall.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications (9th ed.).* Upper Saddle River, NJ: Pearson Education Inc.

Gilfert, S., & Harada, K. (1992). Two composition scoring methods the analytic vs. holistic method. *Hokuriku University Foreign Languages Research Journal, 1,*17-22.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education, 27*(2), 73-82.

Hamp-Lyons, L. (1995). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

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: Atatu̇rk University.

Hinkel, E. (2002). *Second language writers' text: Linguistics and rhetorical features.* Mahwah, NJ: Lawrence Erlbaum Associates.

Homburg, T. J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87-108.

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.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huang, 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. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443.

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

Huang, J., Han, T. (2013). Holistic or analytic – A dilemma for professors to score EFL essays? *Leadership and Policy Quarterly, 2*(1), 1-18.

Han, T., & Ege, İ. (2013). Using generalizability theory to examine classroom instructors' analytic evaluation of EFL writing. *International Journal of Education, 5*(3), 20-35. doi:10.5296/ije.v5i3.3713

Huot, B. (2002). *(Re)Articulating writing assessment: Writing assessment for teaching and learning.* Logan, Utah: Utah State University Press.

Information Technology Services. (2013). *SPSS for Windows: Getting started.* Texas: The University of Texas at Austin. Retrieved on April, 20, 2013 from http://www.utexas.edu/its-archive/rc/tutorials/stat/spss/spss1/

Johns, A. M. (1991). Interpreting an English competency examination: The frustrations of an ESL science student. *Written Communication, 8*(3), 379-401.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: The Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U., Read, J., & Randow, J. V. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20,* 148-161.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computed-delivered writing test. *Assessing Writing, 9*(1), 4-26.

Lee, Y.-W., Gentile, C., & Kantor, R. (2009). Toward automated multi-trait scoring of essays: Investigating links among holistic,

analytic, and text feature scores. *Applied Linguistics, 31*(3), 391-417.

Lenel, J. (1990). The essay examination part III: Grading the essay examination. *The Bar Examiner, 59*(3), 16-23.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing* [Proceeding]. Paper presented at the 3rd annual JALT Pan-SIG Conference. Japan: Tokyo Keizai University.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly,* 17(4), 651-671.

Rezaei, A.R & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing 15,*18–39

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*, 189-209.

Russikoff, K. A. (1995). *A comparison of writing criteria: Any differences?*, [Proceeding]. *Paper presented at the annual meeting of the Teachers of English to Speakers of Other languages*, Long Beach: CA.

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

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.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*, 163-182.

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.

Sudweek, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. B. S. Everitt & D. C. Howell, (Eds.). *Encyclopedia of statistics in behavioral science* (pp.717-719). Chichester: John Wiley & Sons.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11,* 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-87.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

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.

White, E. M. (1994). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.

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 on April 20, 2015 from http://www.eric.ed.gov/PDFS/ED461992.pdf