English Language Teaching; Vol. 8, No. 7; 2015
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Holistic versus Analytic Evaluation of EFL Writing: A Case Study

Thikra K. Ghalib[1] & Abdulghani A. Al-Hattami[2]

[1] Taiz University, Yemen

[2] University of Dammam, Saudi Arabia

Correspondence: Thikra K. Ghalib, Taiz University, Yemen. E-mail: thikrakaed@yahoo.com

Received: April 21, 2015 Accepted: May 23, 2015 Online Published: June 29, 2015

doi:10.5539/elt.v8n7p225 URL: http://dx.doi.org/10.5539/elt.v8n7p225

Abstract

This paper investigates the performance of holistic and analytic scoring rubrics in the context of EFL writing. Specifically, the paper compares EFL students' scores on a writing task using holistic and analytic scoring rubrics. The data for the study was collected from 30 participants attending an English undergraduate program in a Yemeni university. The authors used psychometric statistics (Inter-rater Agreement, Intra-Class Correlation, *t*-test and ANOVA) to compare the performance of the students on the two rubrics in accurately diagnosing students' strengths and weaknesses and placing them along a continuum of foreign language writing proficiency. The raters of the writing samples included three experienced instructors working at the same department. The results of correlating the students and raters' holistic and analytic scores and of examining the variations among the correlations provide evidence for the reliability and validity of both rubrics. Analytic scoring rubrics, however, placed the examinees along a more clearly defined scale of writing ability, and are, therefore, more reliable than holistic scoring rubric instruments for evaluating EFL writing for achievement purposes than holistic scoring rubric.

Keywords: holistic scoring, analytic scoring, EFL writing, writing rubrics, psychometric statistics

1. Introduction

'Good writing' is a growing pedagogic demand. In educational settings, writing is the basis upon which a candidate's achievement, learning and intelligence are judged. Good writing skills are critical to academic and professional success—they can lead to good grades, admission into college, exit from college, a good job, and upward professional movement.

Consequently, and expectedly, testing for writing ability is becoming a very pressing demand. The purposes for which the writing ability is tested include, but are not limited to, awarding grades, certifying proficiency, testing suitability for a particular profession, placing candidates in the appropriate component of a language program, and allowing candidates to exit programs. While the stakes are not high for some of these purposes, they are very high for others—they have important consequences that significantly impact the test taker's life.

To test for writing ability is to define 'good writing'. The measurement of the writing ability is impacted by four factors, namely, the student, the scoring method, the test administration, and the test itself (Mousavi, 2002). While all other three factors are equally significant, the most relevant to the concerns of the present paper is the scoring method—the method selected by the rater to pass judgments about the writing ability. There have been numerous attempts in the literature to introduce methods of scoring (e.g., Hamp-Lyons, 1991; Shohamy, 1995) and many other attempts improve the accuracy and consistency of these methods (McNamara, 1996; Brown, 1996; Wiseman, 2012). The decisions about writing competence that are derived from one scoring method do not always, and do not necessarily, comply with decisions from another scoring method. These scoring systems are very important because they are used to classify test takers and, accordingly, make high-stakes decisions that define the course of their lives.

Moving down to our research context, the success of undergraduate students of English at Taiz University in Yemen is also largely dependent on their ability to write. The program is eight-semester long, comprising 52 courses among which 45 are on English language and literature. Each of the 45 English courses involves a mid-term test and a final exam—a total of 90 achievement tests overall. Passing these achievement tests and exiting the program rests mainly on the students' writing ability. The improvement of the ability to write in English is, therefore, one of the objectives of the program of instruction.

Informal interviews with the teaching staff of the department and an examination of a random sample of mid-term and final exam answer-books suggest that a general-impression marking scheme is the norm. The criteria of evaluation, according to the teacher-raters, are content relevance, content coverage, and language. Test takers who address all the points adequately are rewarded and those who do not are penalized. There are no clear descriptors, however, for awarding marks to intermediate levels of writing proficiency, other than a general impression. The descriptors are not explicitly stated; they are neither clear to the teacher-rater, nor are they known to the test takers. The result is impressionistic judgments of writing proficiency that depend more upon the rater than upon text qualities, and that fail to make valid distinctions between test takers across a continuum of writing proficiency.

In light of these considerations, it becomes of paramount importance to improve consistency across evaluator's judgments about writing proficiency and to improve the reliability and validity of these judgments in order to avoid bias and produce greater agreement between raters about test taker's achievement. An important move towards achieving this objective is using scoring rubrics. Different kinds of rubrics have been in use since, at least, the 1960s and have received much scholarly attention. This paper will focus on the major types of rubrics to measure writing proficiency, considers the uses of each scoring rubric, and outlines the theoretical and practical advantages and disadvantages of each.

2. Literature Review

A rubric is a tool for evaluating the quality of student work on a continuum of performances from excellent to poor (Schafer, 2004). It contains a set of well-established criteria corresponding to a scale of possible points to be assigned in scoring a piece of work, spoken or written (Campbell, Melenyzer, Nettles, & Wyman, 2000). The best performance is assigned the highest point and the worst the lowest point on the scale. A scoring rubric provides descriptors for the different levels of proficiency on the scale. These descriptors are detailed enough to enable sufficiently fine judgments, and rich enough to enable reliable, unbiased and valid discrimination.

Herman, Aschbacher, and Winters (1992) posited four characteristic features of a rubric—criteria, standards, scale, and examples. An effective rubric has a well-defined list of criteria for the test-takers to know what is expected of them and for the raters to be able to properly evaluate the responses. Second, an effective rubric contains standards of excellence for the different levels of performance. Third, an effective rubric has gradations of quality, or a scale, based on the degree to which the standard has been met. The gradations are constituted by detailed descriptions that represent what should earn which point on the scale. Last, but not least, an effective rubric contains modal exemplars of expected performance at the different levels on the scale.

Another important characteristic of a rubric—one that is well attested in the literature though not mentioned in the previous list, is reliability. An effective rubric is one that is used by different raters on a given assessment task and generates similar judgments/scores. Consistency across raters' judgment about the relative standing of performance ratings is referred to as "inter-rater reliability", and the frequency of two or more raters assigning the exact same rating to a particular performance is known as "inter-rater agreement". While these two forms of reliability estimators are frequently employed in research contexts, inter-rater agreement is more relevant to the present research context where decisions about passing exams, exit programs, and even about tenure are made based on a score threshold. In Yemen, for example, a student receiving 47 marks fails the test whereas a student receiving 48 is pushed to the cut-off score and passes the test.

Weigle (2002) argued that there are three types of rubrics used in the evaluation of written proficiency. These are primary trait, holistic and analytic scoring rubrics. These three types differ in their impact, discriminatory power, inter-rater reliability, the degree of bias, and the cost-effectiveness—in terms of time, effort and money (Kuo, 2007). The choice of one scoring rubric or the other is significant because if "represents, implicitly or explicitly, the theoretical basis upon which [a] test is founded" (Weigle, 2002, p. 109). The relevance of this paper, some studies that have used holistic scoring rubrics, analytic scoring rubrics, and studies that have compared both types of rubrics are reviewed below.

Holistic scoring is "a global approach" to scoring that is underscored by the idea that "writing is a single entity which is best captured by a single scale that integrates the inherent qualities of the writing" (Wiseman, 2012, p. 59). As such, holistic scoring considers the entire written response and assigns an overall score to the performance (White, 1985; Weigle, 2002; Hyland, 2002). This cost-effectiveness of holistic scoring makes it a suitable approach for large-scale assessment of written performance, especially for decisions concerning placement (Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993).

Holistic scoring criteria consist of general guidelines that define good performance at each score point. This has prompted a number of researchers (e.g., White, 1985; Cohen, 1994) to argue that holistic scoring focuses on the strengths of the writing rather than on the deficiencies. The holistic rubric generates a composite score that "does not provide specific evidence of where and how much additional writing instruction is needed" (Becker, 2011, p. 116). Despite this shortcoming, if indeed it is, Weigle (2002) argues that holistic scoring rubrics are very practical. They are short, do not include detailed criteria of evaluation, and make possible the evaluation of an essay by assigning one score to it after only one reading—thus serving the economic interests of university departments and employers. Holistic rubrics are therefore typical for evaluating written performance in large-scale assessment contexts. This has made holistic scoring the method of assessing written performance in the computer-based Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), and Graduate Management Admission Test (GMAT).

Diederich (1964) was one of the earliest studies to make use of holistic scoring rubrics in such large-scale testing situations. Three hundred written performances were evaluated by fifty-three raters, and the study concluded that the variation is the ratings is mostly attributable to three criteria—ideas, language and organization. Twenty years later, Breland and Jones (1984) analyzed eight hundred written samples and also attributed the variations of raters to ideas, organization, and use of supporting materials. Successive other studies have examined the issues of the validity of holistic scoring (Charney 1984), inter-rater reliability (Stach, 1987; Erickson, 2001), the consistency of agreement among raters (Huot,1990; Legg, 1998), the importance of rater training for achieving internal consistency and normative rating behavior (Kondo-Brown, 2002; Kim, 2010), the difference in the ratings of native and non-native English speaking raters in China (Shi, 2001), and alternative methods of evaluating writing performance (Reid, 1993).

As an alternative, analytic scoring, which involves "the separation of the various features of a composition into components for scoring purposes", has also received considerable scholarly attention (Wiseman, 2012, p. 60). An analytic scoring rubric typically includes writing components relating to the test taker's lexical, syntactic, discourse, and rhetorical competence. As such, an analytic scoring rubric offers more detailed information about a test taker's writing performance than does the single score of a holistic scoring rubric. An analytic rubric provides orderly and comprehensive feedback to teachers and assists them in the discrimination of the weak and strong aspect in students writing performance (Hamp-Lyons, 1995; Crehan, 1997). In other words, an analytic rubric has higher discriminating power (Mendelsohn & Cumming, 1987).

The first analytic scoring rubric was the ESL Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). It was used to measure the writing performance of ESL students at North American universities and consisted of five different rating dimensions of writing quality, each having a different weight: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). Other well-known examples of analytic scales are the Test in English for Educational Purposes (TEEP; Weir, 1990) and the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991). But of all the existing rating scales for examining written performance (see Shohamy, 1995), the present study adopts, indeed adapts, Bachman and Pamer's (1996) model of communicative language ability and the rubric based on the model. According to the model, the ability to write an essay requires knowledge schemata (knowledge of the topic), strategic competence (strategies for content development), rhetorical knowledge (strategies for producing cohesive supporting arguments), grammatical competence, and knowledge of vocabulary and register. This is the knowledge that defines L2 writing ability in Bachman and Palmer's approach and the knowledge that informs their analytic scoring rubric.

But which scoring rubric, holistic or analytic, is more preferred by practitioners? There are a number of studies that have compared the behavior of holistic and analytic rubrics with interesting findings. Chi (2001) compares holistic and analytic scoring rubrics, using many-faceted Rasch measurement, in terms of the appropriateness of the scoring rubrics, the agreement of the student scores, and the consistency of rater severity. The study reports significant differences between raters using holistic scoring rubrics, but not analytic scoring rubrics. Other studies confirm this advantage of analytic scoring in terms of inter-rater and intra-rater reliability (Al-Fallay, 2000; Easy & Young, 2007; Knoch, 2009; Nakamura, 2004). Analytic scoring also provides an individualized profile of the test taker's written performance (Weigle, 2002) and direct, useful feedback to students and teachers (Brown & Hudson, 2002). For this reason, analytic scoring rubrics are often chosen for placement and diagnostic purposes (Jacobs et al., 1981; Perkins, 1983; Hamp-Lyons, 1991).

By contrast, holistic scoring rubrics offer the advantage of reduced cost in time and money (Wiseman, 2012). Bauer (1981) compared the cost-effectiveness of analytic and holistic scoring rubrics in scoring secondary school students' essays. The study reports that the time needed to train the raters to use the analytic rubrics was two times the time needed to train raters to use the holistic rubrics, and the time needed to grade the essays using the analytic rubrics was four times the time needed to grade the essays using the holistic rubrics. Other studies in different other contexts have reported similar findings (Klein et al., 1998; Arter, 1993; Bainer & Porter, 1992). For this reason, holistic scoring is the preferred method of scoring in large-scale testing contexts that involve a large concentration of test takers taking the test at the same time (Becker, 2011).

The choice of one type or rubric or the other, therefore, depends mainly on the purpose of using the rubric and is driven by context-specific considerations. The present study is an extension of this tradition of examining the performance of holistic and analytic scoring rubrics. The study used different psychometric statistics (Inter-rater Agreement, Intra-Class Correlation, *t*-test and ANOVA) to compare the holistic and analytic scoring rubrics as reliable instruments for evaluating EFL writing for achievement purposes. The authors of this study tried to find answers to the following question:

1) Is there a significant difference between holistic and analytical rubrics in enhancing the reliability of scoring?

2) Is there a correlation between each rater's assessment of the same essay using holistic and analytic rubrics?

3) Is there a correlation between different raters' assessment of the same essay using holistic rubrics?

4) Does the use of rubrics enhance the consistency of scoring?

3. Methods

3.1 Participants

The participants of the study consisted of 30 male and female Yemeni undergraduate students of English at the Faculty of Arts, Taiz University. They were aged between 21 and 25, and were all non-native speakers of English attending the three-credit, 14-week senior-level course Advanced Writing Skills. The course is offered in the seventh semester of the eight-semester Bachelor Program in English Language and Literature. The researcher chose the participants on the basis of their overall GPA in the first three years of college. The participants in this study were the top 30 students in the six semesters leading to the year 2014-2015. They were the senior students of the English department. They took a class on advanced writing skills and have reached a level of competence that should enable them to write an essay. Therefore, the authors wanted to get the most competent students in terms of merit.

3.2 Raters

The raters of the writing task consisted of three experienced teaching staff of the same department. They were selected based on their similarity in terms of qualifications, years of experience in teaching, and years of experience in scoring high-stakes tests. The three raters all had a doctoral degree in English with at least five years of teaching experience at the same department. They also had taught different writing courses at the department, and had marked at least three rounds of the annual large-scale English admission test administered by the department.

The raters were invited to a two-hour training session conducted by the researchers. The training, which eventually aimed at improving rating accuracy and rater agreement, involved an explanation of the rating system, a discussion of common rating problems, and advice on avoiding bias.

3.3 Scripts and the Writing Task

The scripts consisted of essays written by the 30 participants in response to an independent, timed writing task. The task prompt to the essay was as follows:

Reflecting on your own first day in college, write a descriptive essay of about 250 words in response to the following question, *What was your first day in college like? How did you feel as a new comer? And what did you do?*

3.4 Rating Rubrics

The study employed two rubrics—a holistic rubric and an analytic rubric. The holistic rubric is a six-point scale that offers a general description at each point for typical writing performance at that point (see Appendix A). It emulates the rubric used by teachers of the department for assessing students' performance on written tasks. In fact, it has been constructed by the researchers after informal interviews with the teachers about the criteria they use for evaluating written work. The suggested rubric, therefore, comprises two performance criteria—understanding of the topic and correctness of language.

The analytic rubric, on the other hand, is an adapted version of Bachman and Palmer (1996). The researchers contributed a fifth sub-domain to Bachman and Palmer's criterion-referenced rating scale for the assessment of writing ability. This addition was driven by context-specific considerations. The end product is a five-point scale with five sub-domains of writing ability, viz., content, cohesion, syntactic structures, vocabulary, and mechanics of writing. Within each domain, there are several well-defined standards of performance points that each rater clearly understands (see Appendix B).

3.5 Rating Procedures

Each rater worked independently and in two separate sessions. In the first session, the raters were given the 30 (anonymous) writing samples and a copy of the holistic rubric. The raters were instructed to assign a single 'holistic' score to each essay from 0 to 5. The scores were then converted into 20 and the total score written next to the number assigned to each participant. The scored writing samples and the rubrics were returned to the researchers in three days' time. The second session took place a month later to allow a gap long enough to ensure a more independent judgment. In this session, the raters were given the same 30 (also anonymous) writing samples and copy of the analytic rubric. They were instructed to assign a score from 0 (zero knowledge) to 4 (complete knowledge) for each sub-domains of writing proficiency and then add the scores and convert them into a total of 20. The scored writing samples and the rating rubrics were returned to the researchers in a week's time.

4. Results

A number of statistical procedures were employed to answer the study research questions. First, the descriptive statistics of the students' scores using the holistic and analytic rubrics were calculated. This was followed by the descriptive statistics of each rater's assessment of the writing sample using both rubrics. A *t*-test was used to examine if there was a significant difference between the means of the two scoring rubrics, and Analysis of Variance was conducted to examine if there were any significant differences between the three raters' scoring decisions for each of the two scoring rubrics. In addition, to investigate the agreement among the three raters and measure the inter-rater reliability, an Intra-Class Correlation Coefficient test was implemented. The findings of this study are discussed below.

The results showed that the mean score of the scores using holistic rubric was 14.67 with a standard deviation of 3.12. Using the analytical rubric to assess students' performance yielded a mean of 13.72 and a standard deviation of 2.82.

Table 1. Descriptive statistics for each of the scoring rubrics (N = 90)

Rubrics	Minimum	Maximum	Mean	Std. Deviation
Holistic	8	20	14.67	3.116
Analytical	6	19	13.72	2.821

Descriptive statistics for each of the three raters within each of the two rubrics are presented below.

Table 2. Descriptive statistics for each of the three raters within each of the two rubrics (N = 30)

Rubrics	Raters	Minimum	Maximum	Mean	Std. Deviation
Holistic	Rater # 1	8	20	14.67	3.536
	Rater # 2	12	20	15.87	2.675
	Rater # 3	8	16	13.47	2.675
Analytical	Rater # 1	6	18	13.73	3.162
	Rater # 2	9	19	14.03	2.606
	Rater # 3	6	18	13.40	2.724

A *t*-test was performed to examine if there was a significant difference between the means of the two scoring rubrics, holistic and analytic. The results showed that the difference was significant between the two rubrics, $t_{(178)} = 2.132$, $p < .05$. Using the analytical rubric proved to be more rigorous ($M = 13.72$, $SD = 2.821$) than using the holistic approach of scoring ($M = 14.67$, $SD = 3.116$).

Assessment should be independent of who does the scoring and the results are supposed to be similar. The more consistent the scores are over different raters, the more reliable the assessment is. Analysis of Variance was used to investigate if there were any significant differences between the three raters for each of the rubric method. The findings showed that there were no significant difference, $F_{(2, 87)} = 0.373$, $p = 0.690$, among the three raters when they used analytical rubric to grade students' performance. However, the raters scorings did significantly differ, $F_{(2, 87)} = 4.833$, $p < .05$, when they used holistic rubric. Post Hoc analysis was run to find where the differences lie. The results showed that the difference was between rater 2 and rater 3 at $P < 0.05$.

It is worth investigating to check the correlation between the two scoring methods. If the correlation is high, that means that the two scoring methods may produce similar results. The results here indicated that there was a highly significant correlation, $r = 0.80$. Nevertheless, a correlation in this context should be more than 0.90.

Studies in literature indicated that rubrics seem to aid raters in achieving high internal consistency when scoring performance tasks. Intra-Class Correlation Coefficient was used to measure intra-rater reliability, the average measures equals the reliability across the raters. For Holistic Rubric, the average measure of ICC for the holistic rubric was .797 with a 95% confidence interval from .567 to .904 ($F_{(29, 58)} = 6.627$, $p <.001$). Whereas, for Analytical Rubric, the average measure of ICC for the analytical rubric was .958 with a 95% confidence interval from .921 to .979 ($F_{(29, 58)} = 25.364$, $p <.001$). Overall, a high degree of reliability was found for the internal consistency. The average measure of ICC was .879 with a 95% confidence interval from .788 to .930 ($F_{(59, 118)} = 10.104$, $p <.001$).

Cohen's kappa was also used to estimate the degree to which there is an agreement among the raters. The results for each pair of the three raters and the overall across the two scoring rubrics are presented in Table 3.

Table 3. Descriptive statistics for each pair of the three raters and the overall across the two scoring rubrics (N = 30)

	Pair 1 & 2	Pair 1 & 3	Pair 2 & 3	Total Reliability
Overall	.51	.25	.09	.28
Holistic	.49	.14	-.05	.23
Analytic	.42	.18	.04	.21

5. Discussion

Assessment of students performance has to be as accurate as possible because it may have consequences for students being assessed (Black, 1998). There are some sources of variability in any assessment, one of which is raters' judgments of students' performance (Black, 1998). This was the focus of this study.

The difference between holistic and analytical rubrics in enhancing the reliability of scoring was investigated and the results of this study showed that there was a significant difference between the means of the two scoring rubrics, holistic and analytic approaches. It was found that when raters use analytical scoring rubric, they give lower scores than when using holistic scoring rubric. Such findings make sense because analytical rubrics have many details and scoring them is more rigorous. Studies in literature indicated that for this reason analytical scoring rubrics are often used for diagnostic purposes (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981; Perkins, 1983; Hamp-Lyons, 1991).

The correlation between the two scoring methods was also computed. The results showed that there was a highly significant correlation. However, this does not mean that there is an agreement among the raters; another analysis was conducted below. The correlation between two scoring methods was 0.80 which is deemed acceptable (Stemler, 2004).

The students' scores are supposed to be similar regardless of who does the scoring. The more consistent the scores are over different raters, the more reliable the assessment is. The Analysis of Variance (ANOVA) showed that there were no significant differences among the three raters when they used analytical rubric to grade students' performance. However, the raters' scorings did significantly differ when they used holistic rubric. These findings are consistent with Chi (2001) findings about the significant differences between raters using holistic and analytical scoring rubrics. The more consistent the scores are over different raters, the more reliable the assessment is thought to be (Moskal & Leydens, 2000). The findings in this study suggest that using analytical rubric produce more consistent and reliable results.

Variations in raters' judgments can occur either across raters, known as inter-rater reliability, or in the consistency of one single rater, called intra-rater reliability. Intra-class Correlation was performed to measure interrater reliability and the consistency of the raters in measuring the students' performance. The Intra-Class Correlation Coefficient was above 0.80 indicating that results are consistent. The majority of studies investigating intra-rater reliability reported alpha values above 0.70 which, according to Brown, Glasswell, and Harland (2004), is generally considered sufficient.

An interrater agreement refers to the extent to which independent raters provide the same rating of a particular person. Cohen's kappa was used to estimate the degree to which consensus agreement ratings vary from the rate expected by chance. The results of study showed that the correlation between two raters appears to be high, and the correlation between two other raters appeared to be low. Kappa values between 0.40 and 0.75 represent fair agreement beyond chance (Stoddart, Abrams, Gasper, & Canaday, 2000)

6. Conclusion

Rubrics are used by teachers to evaluate students' performance on specific tasks. A rubric is a scoring scale used to assess students' performance along a task-specific set of well-defined criteria. A number of benefits were discussed for using rubrics as a tool to evaluate students on performance tasks. The use of rubrics can 1) increase the consistency of judgment when assessing performance tasks, 2) provide is a valid judgment of performance assessment that cannot be achieved by not using the rubric, 3) give positive educational consequences, such as promoting learning and/or improve instruction, and 4) provide students with quality feedback (Jonsson & Svingby, 2007; Archbald & Newmann, 1988).

Having explored the differences between the widely-used scoring systems for wiring ability, and having underscored the importance of implementing rubrics for better diagnosis of writing problems and for more reliable scoring, the present study zooms in on two kinds of rubrics, viz., holistic and analytic rubrics, and examines the performance of these two rubrics in assessing writing ability. Specifically, the study compares Yemeni EFL students' scores on a writing task using holistic and analytic scoring rubrics.

This study analyzed different psychometric statistics to compare the holistic and analytic scoring rubrics as reliable instruments for evaluating EFL writing for achievement purposes. The results showed that using rubrics yields more accurate scores than not using them; this was also clearly stated in the literature. However, it was concluded that analytical scoring provides even more consistent scores than using holistic scoring methods. Analytical scoring seems to be very useful in the classroom because the results can help both the teachers and learners identify students'' strengths and weaknesses as well as the learning needs. As educators we need to accept that the use of rubrics add to the quality of the assessment (Perlman, 2003).

In summary, scoring with rubrics seems to be more reliable than scoring without one. Rubrics ought to be encouraged as a regulatory device for scoring. The results in this study showed that using holistic rubric can give reliable scores and using analytic rubric gives even more reliable scores. The consistency of scoring can be enhanced by being analytic, topic-specific, and rater training.

7. Limitation

The main aim of this paper was measuring the consistency of scoring across raters' judgment by means of different correlation coefficients using Many-Facets Rasch Model (MFRM). However, due to the small sample size, MFRM was not used in this study. MFRM is a multivariate extension of Rasch measurement models that can be used to provide a framework for calibrating both raters and writing tasks within the context of writing assessment. Another limitation is that the study was conducted in one institution and used a convenience sample. Therefore, we recommend using a larger and random sample of students from multiple institutions for future research.

References

Al-Fallay, I. (2000). Examining the analytic marking method: Developing and using an analytic scoring schema. *Language & Translation*, *12*, 1-22.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond Standardized Tests*. Reston, Va.: National Association of Secondary School Principals.

Arter, J. (1993, April). Designing scoring rubrics for performance assessments: The heart of the matter. Paper presented at *the annual meeting of the American Educational Research Association, Atlanta, GA*.

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bainer, D., & Porter, F. (1992, October). Teacher concerns with the implementation of holistic Scoring. Paper presented at *the annual meeting of the Midwestern Educational Research Association, Chicago*.

Bauer, B. A. (1981). *A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability*. Champaign: University of Illinois.

Becker, A. (2011). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal*, *22*(1), 113-130.

Black, P. (1998). *Testing: Friend or foe*? London: Falmer Press.

Breland, H. M., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, *1*, 101-109.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*, 105-121.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Campbell, D. M., Melenyzer, B. J., Nettles, D. H., & Wyman, R. M. Jr. (2000). *Portfolio and performance assessment in teacher education*. Boston: Allyn and Bacon.

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

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement*, *2*(4), 379-388.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston, MA: Heinle & Heinle.

Crehan, K. (1997, October). A discussion of analytic scoring for writing performance assessments. Paper presented at *the Arizona Education Research Association, Phoenix, AZ*.

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

Diederich, P. B. (1964). Problems and possibilities of research in the teaching of written composition. In D. H. Russell, M. J. Early, & E. J. Farrell (Eds.), *Research Design and the Teaching of English: Proceedings of the San Francisco Conference 1963*. National Council of Teachers of English, Champaign, IL, 52-73.

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics*, *13*, 1-21.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). New York: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL context. In L. Hamp-Lyons (Ed.), *Assessing second language writing* (pp. 241-277). Norwood, NJ: Ablex.

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

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Huot, B. A. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-211.

Hyland, K. (2002). *Teaching and research Writing*. Harlow, England: Pearson Education Limited.

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

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

Kim, H. J. (2010). *Investigating raters' development of rating ability on a second language speaking assessment* (Unpublished doctoral dissertation). Teachers College, Columbia University.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*, 121-137.

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Kuo, S. (2007). Which rubric is more suitable for NSS liberal studies? Analytic or holistic? *Educational Research Journal*, *22*(2), 179-199.

Legg, S. M. (1998). Reliability and Validity. In W. Wolcott, & S. M. Legg (Eds.), *An overview of writing assessment: Theory, research, and practice* (pp. 124-142). Urbana, Ill: National Council of Teachers of English.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal*, *5*(1), 9-26.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*, 71-81.

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed.). Taipei: Tung Hua Publications.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing*. Retrieved April 9, 2015, from https://jalt.org/pansig/2004/HTML/Nakamura.htm

Panou, D. (2013). L@ writing assessment in the Greek school of foreign languages. *Journal of Language Teaching and Research*, *4*(4), 649-654. http://dx.doi.org/10.4304/jltr.4.4.649-654

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

Reid, J. (1993). *Teaching ESL Writing*. Englewood Cliffs, NJ: Regents Prentice Hall.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18-39.

Schafer, L. (2004). *Rubric*. Retrieved February 9, 2015, from http://www.etc.edu.cn/eet/articles/rubrics/index.htm

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303-325.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4).

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning—A report of methodology. *International Journal of Science Education*, *22*, 1221-1246.

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

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall Regents.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Wiseman, S. (2012). A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing. *Iranian Journal of Language Testing*, *2*(1), 59-92.

Appendix A

Holistic Rubric

5) Provides a complete response to the prompt; demonstrates complete understanding of the topic; exhibits a strong command of essay writing skills; presents argument in flawless English

4) Provides a fairly complete response to the prompt; demonstrates considerable understanding of the topic; exhibits good knowledge of essay writing skills; presents argument in very good English with a few errors

3) Provides a satisfactory response to the prompt; demonstrates partial understanding of the topic; exhibits a limited command of essay writing skills; presents argument in good English with many errors

2) Provides a poor response to the prompt; demonstrates little understanding of the topic; exhibits little command of essay writing skills; presents argument in poor English with too many errors

1) Provides no poor response to the prompt; demonstrates no understanding of the topic; exhibits no command of essay writing skills; presents argument in barely comprehensible English

0) No response

Appendix B

Analytical Rubric

Mechanics of writing	Range: no evidence of knowledge of the mechanics of writing Accuracy: not relevant	Range: little evidence of deliberate use of correct spelling, punctuation, capitalization and paragraphing techniques Accuracy: poor or moderate accuracy
Rating	0. Zero Knowledge	1. Limited Knowledge
Vocabulary	Range: no evidence of knowledge of vocabulary—inadequate even of simplest formal vocabulary Accuracy: not relevant	Range: small range lacking the formal and appropriate vocabulary required to produce good piece of writing. Accuracy: vocabulary items frequently used imprecisely (limited success in conveying meaning)
Rating	0. Zero Knowledge	1. Limited Knowledge
Syntactic structures	Range: no evidence of knowledge of syntactic structures—inadequate to use even the simplest structures Accuracy: not relevant	Range: small range including a few basic structures Accuracy: poor to moderate accuracy within range; if structures outside of the controlled range are attempted, accuracy may be poor
Rating	0. Zero Knowledge	1. Limited Knowledge
Cohesion	Range: no evidence of knowledge of cohesive relationships Accuracy: not relevant	Range: a few markers of cohesion Accuracy: relationships between sentences frequently confusing; composition barely intelligible
Rating	0. Zero Knowledge	1. Limited Knowledge
Content	Range: inadequate to produce even the simplest organized text Accuracy: not relevant	Range: little evidence of deliberate, correct and relevant text Accuracy: organization generally unclear or irrelevant to topic
Rating	0. Zero Knowledge	1. Limited Knowledge

Range: moderate range of proper spelling, punctuation, capitalization and paragraphing techniques Accuracy: moderate to good accuracy but could be more explicitly marked	Range: wide range of proper spelling, punctuation, capitalization and paragraphing techniques Accuracy: good accuracy, few errors but these errors do not affect the meaning that is conveyed accurately
2. Moderate Knowledge	3. Extensive Knowledge
Range: moderate range—sufficient to produce a fairly comprehensible piece of writing Accuracy: vocabulary items sometimes used imprecisely (some paraphrasing is used)	Range: wide range of general and specific vocabulary Accuracy: vocabulary items adequately cover the assigned task and are seldom used imprecisely
2. Moderate Knowledge	3. Extensive Knowledge
Range: medium range—uses basic structures and avoids complex structures Accuracy: moderate to good accuracy within range; if structures outside of the controlled range are attempted, accuracy may be poor	Range: wide range basic structures with some uses of complex structures Accuracy: good accuracy, few errors but these errors do not affect the meaning that is conveyed accurately
2. Moderate Knowledge	3. Extensive Knowledge
Range: moderate range of explicit textual devices Accuracy: relationships between sentences generally clear but could often be more explicitly marked and the composition could be more fluid and intelligible	Range: wide range of explicit cohesive devices including complex subordination Accuracy: highly accurate with only occasional errors_in cohesion; composition easily intelligible
2. Moderate Knowledge	3. Extensive Knowledge
Range: moderate range of explicit text organizational devices Accuracy: organization generally clear but could often be more explicitly marked	Range: wide range of explicit text organizational devices on essay and paragraph levels Accuracy: highly organized text
2. Moderate Knowledge	3. Extensive Knowledge

Range: evidence of unlimited range of proper spelling, punctuation, capitalization and paragraphing techniques; Accuracy: evidence of complete accuracy of use	
4. Complete Knowledge	
Range: evidence of complete range of vocabulary Accuracy: evidence of complete accuracy of usage	
4. Complete Knowledge	
Range: evidence of unlimited range; using syntactic structures ranging from simple to complex Accuracy: evidence of complete control or accuracy	
4. Complete Knowledge	
Range: evidence of complete range of cohesive devices Accuracy: evidence of complete accuracy of use; composition perfectly intelligible	
4. Complete Knowledge	
Range: evidence of complete range of explicit text organizational devices Accuracy: evidence of complete accuracy of use	
4. Complete Knowledge	Score

Copyrights