Holistic vs. Analytic Evaluation in Writing Test of Eighth Grade Students

Elda Selja Putri a *, Melyann Melani b

a English Education Study Program, Faculty of Teachers Training State Islamic University,
b Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi, Indonesia

Abstract

The purpose of this study is to explore the efficacy of holistic and analytic scoring rubrics when applied to the context of writing examinations. Throughout this research, a technique was utilized to compare two grading models based on (a) assessor agreement (reliability) and (b) reasons for the grades given (validity). Teachers were randomly allocated to one of two conditions, where they graded the same student performance using either an analytical or a holistic approach. The design is experimental. Agreement and rank correlation between grades have been compared. A comparison of the students' performance on a writing assignment is made in this study using both holistic and analytic scoring rubrics. The information for this study was gathered from twenty-four students in eighth grade at SMP N 5 Kecamatan Kapur IX. The authors conducted an analysis of the performance of the students on the two rubrics, using psychometric statistics, in order to precisely determine the strengths and weaknesses of the students, and to place them along a continuum of foreign language writing ability. Citation In contrast, analytic scoring rubrics placed examinees on a scale of writing competency that was more specifically defined. As a result, analytic scoring rubrics were more dependable than holistic scoring rubrics when it came to gauging writing for the purpose of achieving goals.

© 2022 JELS and the Authors - Published by JELS.
As a direct consequence of this, and as was to be anticipated, the importance of testing writing skill is growing. The objective of writing ability assessment is, among other things, to offer grades, certify competency, evaluate acceptability for a particular occupation, position applicants in the proper portion of a language program, and allow candidates to withdraw from programs. Even if the stakes aren’t as high for some of these objectives, there are others that have far-reaching repercussions that will have a substantial impact on the test-life.

Before evaluating a writer’s skills, it is necessary to first establish what exactly constitutes “excellent writing.” The measuring of a student’s writing skill is affected by a number of factors, including the student, the scoring technique, the test administrator, and the test itself (Santos, 2014) (Mousavi, 2002). The scoring technique, which is defined as “the method used by the rater to evaluate writing talent,” is the component that is most relevant to the problems that are investigated in this study, despite the fact that the other three aspects are each of equal importance. There have been a lot of different attempts made to suggest scoring systems (Hamp-Lyons, 1991), (Shohamy, 1995), and to improve the accuracy and consistency of these different methods (Clapham & McNamara, 2000), (Brown, 1996), and (Wiseman, 2012). It is possible, although by no means certain, that decisions regarding a person’s writing skill that are based on one scoring system will not correlate to those that are based on a different scoring system. These techniques of scoring are necessary because they are used to classify test takers and, as a consequence of this classification, to make judgments that can profoundly affect people’s lives.

To return to the topic of our inquiry, the ability to write is one of the most significant talents that kids in the eighth grade should have. The two English teachers at SMP N 5 Kecamatan Kapur IX need to pass both the midterm and the final examinations in order for them to be able to continue working there. Ability in writing is the single most essential criterion that determines whether or not a student is successful on these achievement assessments and passes on to the next grade level.

At the very least, preliminary exchanges with the teaching staff of the department and an analysis of a random sample of midterm and final exam answer booklets indicate that a general impression marking system appears to be the standard. It is suggested that the criteria for evaluation include the extent to which the subject matter is pertinent, the breadth of the themes covered, and the
language that is employed. Those who answer each question correctly will receive a prize, whilst those who do not will be penalized for their performance. The only thing that can be said with certainty about those with intermediate levels of writing proficiency is that there is a broad perspective on the subject. There are no precise guidelines for assigning grades to these individuals. Because the adjectives are not communicated in a way that is readily apparent, neither the teacher who is evaluating the applicants nor the people who are taking the exam are aware that they are included. As a direct result of this, impressionistic writing competency evaluations lay a greater focus on the rater than they do on the quality of the text, and as a result, they are unable to discern effectively between test-takers along a continuum of writing proficiency.

In light of these considerations, it is of the utmost importance to improve the consistency, reliability, and validity of the judgments that evaluators make regarding the writing proficiency of test-takers in order to reduce bias and promote greater agreement among raters regarding the performance of test-takers. The development and use of grading rubrics is crucial to the achievement of success in this endeavor. Grading rubrics have been used in schools and they have also gained a large amount of academic attention. In this article, the many different types of scoring rubrics that are used to evaluate a writer's writing skills are discussed, along with their applications, as well as the theoretical and practical benefits and drawbacks of using them. Additionally, the article explores the benefits and drawbacks of using these scoring rubrics.

The quality of the student's work is evaluated using a tool known as a rubric, which assigns points ranging from outstanding to unsatisfactory (Schafer, 2004). It includes a list of criteria that have been clearly described and that correspond to a scale of potential points that can be used to grade oral or written work. This scale can be used to grade the work that has been submitted (Campbell et al, 2000). In the event that you have the best performance, you will be awarded the highest possible score, and in the event that you have the poorest performance, you will be awarded the lowest possible score. A grading rubric is used to assist describe the various degrees of expertise that can be found along the spectrum. These descriptions are precise enough to allow for genuine, fair, and accurate discrimination; in addition, they permit the possibility of fine distinctions.
Criteria, standards, a scale, and examples are the four components that Herman, Aschbacher, and Winters describe as constituting a rubric (1992). A successful rubric will have a list of clearly articulated criteria. This will allow test takers to comprehend what is expected of them, and it will also allow raters to assess the findings in the appropriate context. The inclusion of performance standards for a variety of accomplishment levels is the second quality of a useful grading rubric. Thirdly, an efficient rubric will include a scale or quality ratings that are dependent on the degree to which the criterion has been met. The gradations are comprised of in-depth explanations of the factors that should be considered for each point on the scale. Last but not least, a successful grading system will include examples of the modal performance that will be expected at various scale levels.

The reliability of a rubric is another significant trait that is thoroughly established in the literature but was not mentioned in the previous list. This is one of the few aspects of a rubric that has been overlooked. When utilized by numerous raters on the same assessment task, effective rubrics yield judgments and scores that are consistent with one another. The degree to which different raters come to the same conclusion about a performance is measured by something called inter-rater agreement. Inter-rater reliability refers to the frequency with which two or more raters come to the same conclusion regarding a performance and give that conclusion the same rating. Although both types of reliability estimators are frequently used in research, inter-rater agreement is considered to be of greater significance in the modern research context. This is because decisions regarding passing tests, leaving programs, and even tenure are all dependent on a score threshold.

According to Weigle (2002), there are three different kinds of rubrics that are used to evaluate a writer’s proficiency in writing. There are primary trait scoring rubrics, holistic scoring rubrics, and analytical scoring rubrics. The degree to which these three categories are biased, as well as their impact, discriminatory power, inter-rater reliability, and cost-effectiveness (in terms of time, effort, and money), can vary (Kuo, 2007). It is important to choose one scoring rubric over another because “implicitly or openly indicating the theoretical underpinning upon which [a] exam is developed” is one of the reasons that this choice is vital (Knoch, 2009)(Weigle, 2002). There is a discussion of the significance of this study, as well as research that has utilized holistic scoring rubrics, analytic scoring rubrics, and studies that have
contrasted both types of scoring rubrics. Both types of scoring rubrics are compared to one another.

The foundation of holistic scoring, also known as "a global approach" to scoring, is the notion that "writing is a single entity best described by a single scale that incorporates the essential aspects of the writing." This is one of the central tenets of the holistic scoring system (Wiseman, 2012). As a result, holistic scoring takes into consideration the entire written response before determining the final mark for a performance (White, 1985; Weigle, 2002; Hyland, 2002). Because of its low cost, holistic scoring is a good choice for conducting large-scale writing evaluations, particularly when it comes to making placement judgments (Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993). Holistic scoring is a useful option for extensive assessments of writing performance due to its cost effectiveness and cost efficiency.

The global scoring criteria are comprised of standard metrics that, when met, indicated excellent performance at each score point. As a consequence of this, a number of studies (such as White's from 1985 and Cohen's from 1994), suggest that holistic grading lays more attention on the writing's strengths than it does on its shortcomings. According to the grading rubric, the holistic criteria result in a composite grade that "does not provide clear indicators of where and how much extra writing instruction is required" (Becker, 201). Weigle (2002) says that holistic grading rubrics are extremely effective despite the fact that they may have this potential drawback. They are succinct, don't have any specific evaluation criteria, and enable the examination of an essay by assigning a single score after only one reading. As a result, they bring financial benefits to university departments and organizations. It is common practice to use comprehensive rubrics when conducting assessments on a broad scale. These evaluations focus on written performance. As a consequence of this, holistic scoring is utilized in the computer-based Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), and Graduate Management Admission Test in order to measure candidates' ability to write (GMAT).

One of the first studies to use holistic scoring rubrics in such long testing scenarios was Diederich's, which was also one of the earliest studies (1964). According to the findings of the study, the vast majority of the differences in opinions can be traced back to three distinct aspects: concepts, language, and organization. The study found that the majority of the variation in ratings may be attributed to three factors: concepts, language, and organization. Three hundred
written performances were evaluated by fifty-three raters. Breland and Jones (1984) performed an analysis of 800 written examples twenty years later. They ascribed rater disparities to concepts, organization, and the usage of supporting materials. Other research has examined the validity of holistic scoring (Charney, 2009), inter-rater reliability (Stach, 1987; Erickson, 2001), the consistency of rater agreement (Huot, 1990); (Newmark, 1998) (Legg, 1998), the importance of rater training for achieving internal consistency and normative rating behavior (Kondo-Brown, 2002); (Kim, 2011), the difference in the ratings of native and non-native English speaking raters in China (Shi, 2001), and alternative methods (Kondo-Brown, 2002) (Reid, 1993). One of the early research to apply comprehensive grading criteria in lengthy testing scenarios. The research’s conclusions indicate that concepts, language, and organization account for the vast majority of the disparities in beliefs.

In recent years, academics have shown a great deal of interest in analytical scoring, which can be defined as “the partitioning of the many aspects of a piece into components for the purposes of scoring” (Wiseman, 2012). It is a frequent practice to include in a scoring rubric for analytical reasons writing components that are relevant to the test taker’s lexical, syntactic, discourse, and rhetorical abilities. This is done in order to assess the quality of the writing. Because of this, the single score that is obtained from an analytic scoring rubric contains more information about the performance of the test taker in terms of writing than the single score that is derived from a holistic scoring rubric. The use of an analytical rubric provides educators with an organized and thorough review, which in turn assists them in recognizing the areas of their students’ writing abilities that require improvement as well as those that are already strong (Hamp-Lyons, 1995), (Crehan, 1997). To phrase it another way, an analytical rubric has a greater capability for differentiation than a descriptive one (Mendelsohn & Cumming, 1987). Analytical scoring has garnered a lot of interest from academics. Writing elements that are pertinent to the test-lexical, taker’s syntactic, discourse, and rhetorical skills are frequently included in scoring rubrics. To evaluate the writing’s quality, this is done.

The English Language Learner Composition Profile, or ESL Composition Profile to give it its full name, was the first analytical scoring rubric. Its full name was the English Language Learner Composition Profile (Jacobs, et al, 1981). was composed of five distinct rating aspects of writing quality, each of which carried a different weight: content (worth 30 points), organization (worth 20 points), vocabulary (worth
20 points), language use (worth 25 points), and mechanics. Each of these aspects of writing quality carried a different weight. The overall rating was determined by taking the average of the five individual rating components. It was employed as a tool to evaluate the level of writing ability possessed by ESL students who were enrolled in universities located in North American countries. It had five distinct rating dimensions, each of which carried a varying amount of weight, and it had five distinct rating dimensions. The ranking was based on the quality of the writing. The following are the factors that were considered in the grading process: The following aspects each have an equal impact on the final score: the topic (which is worth 30 points), the organization (which is worth 20 points), the vocabulary (which is worth 20 points), the language use (which is worth 25 points), and the mechanics (5 points). Other well-known examples of analytical scales include the Test in English for Educational Purposes (TEEP; Weir, 1990) and the Michigan Writing Assessment Scoring Guide. Both of these tests were developed by Weir. Weir was the one who came up with both of these tests in the first place (Hamp-Lyons, 1991). The model of communicative language competence created by Bachman and Palmer (1996) is the one that is employed and adapted in the current study. The rubric that is based on the model is also the one that is utilized and adapted in this study. The model of communicative language competence that Bachman and Palmer (1996) established is the one that is utilized to grade the writing performance of students today out of all of the grading scales that are now available for evaluating writing performance (see Shohamy, 1995). According to the model, in order to have the ability to write an essay, an individual needs to have knowledge schemata (subject knowledge), strategic competence (techniques for content creation), rhetorical expertise (strategies for building unified supporting arguments), grammatical competence, vocabulary knowledge, and register knowledge. In order to be able to write an essay, you absolutely need to have all of these talents. This is the knowledge that serves as the foundation for Bachman and Palmer’s analytic grading standards, as well as the knowledge that is used by Bachman and Palmer to determine L2 writing competency.

However, more commonly do practitioners use one of the two distinct kinds of scoring rubrics analytical one or a holistic one? The effectiveness of holistic and analytic rubrics has been examined and contrasted in a sizable quantity of study, and as a consequence, the research has yielded some surprising results. (Shi, 2001)
comparing holistic and analytical scoring rubrics in terms of the applicability of the scoring rubrics, the consistency of rater severity, and the agreement of the student scores using the multifaceted Rasch measurement. Regarding the following factors, a comparison is made: the adequacy of the scoring rubrics, the agreement of the student scores, and the consistency of the harshness of the raters. The following factors are taken into account while doing this comparison: the suitability of the scoring rubrics, the consistency of the student scores, and the harshness of the raters' assessments as a whole. The appropriateness of the scoring criteria, the agreement of the student scores, and the consistency of the raters' severity are all taken into account while making this comparison. The results of the study showed that raters who used holistic scoring rubrics and those who used analytical scoring rubrics were very different from one another. Numerous studies have found that analytical scoring improves consistency between and within raters. (Al-ghazo, 2021), (Kuo, 2007)

Analytic scoring offers specific information about the test-taking skills of the person being assessed, as well as rapid and meaningful feedback for students and teachers (Eckes et al., 2016). As a result, analytical scoring rubrics are typically used to make judgments on diagnosis and placement (Jacobs et al., 1981; Perkins, 1983; Hamp-Lyons, 1991). It has been compared and evaluated how well holistic and analytical grading rubrics work. Analytic scoring provides quick and insightful feedback for students and teachers, as well as detailed information about the test-taking abilities of the person being evaluated. Analytical scoring enhances consistency between and between raters, according to recent studies.

However, the advantage of holistic scoring rubrics is that they minimize the amount of time and money lost throughout the grading process (Wiseman, 2012). When analyzing the papers that were submitted by secondary school students, Bauer (1981) contrasted the cost-effectiveness of using an analytical scoring rubric to that of a holistic scoring rubric. It was shown that the holistic scoring rubric would be the most expensive. According to the study's conclusions, it took almost twice as long to train raters on how to utilize analytical rubrics as it did to teach them how to use holistic rubrics. The time needed to assess essays using analytical rubrics was also four times longer than the time needed to evaluate essays using holistic rubrics. Research conducted under various conditions. When there is a need for large-scale testing with many test-takers completing the exam concurrently, the holistic scoring system is the one that is advised for usage. Because of this, it is advised to adopt the holistic
scoring approach when evaluating performance in certain situations. To be more explicit, it is advised that this scoring method be applied in situations like these.

The decision between one type of rubric and another is often based on the rubric’s aim and is impacted by elements that are particular to the setting in which the rubric will be used. In other words, a choice is made to use a specific sort of rubric rather than another. The present investigation continues a decades-long line of investigation on the effectiveness of holistic and analytical scoring rubrics. The current study has examined this area of study. In this line of study, the usefulness of the two various types of scoring rubrics has been examined. A range of psychometric data were used to assess the validity of holistic and analytical scoring rubrics for use in grading English as a foreign language (EFL) writing assignment. This was done to ascertain which of the two methods was more precise. Where the research was done was in the United Kingdom.

**METHOD**

This study’s design concept is experimental, and several professors were split into the analytic and holistic grading conditions. Teachers from several schools in the same region take part in the study. Only grades and written explanations from the teachers have been collected; no personal information.

**Participants**

Eighth Grade Students of SMP N 5 Kecamatan Kapur IX took part in the study. They were all non-native English speakers. The classes are part of second semester of 2021/2022 academic year. The participants were chosen based on their ability during their first at eighth grader. There were three English teachers, there were taught by same syllabus, curriculum, material and time allocation. The students of both classes had similar ability in English subject. To know their homogenous, researcher had done homogenity test to students first semester test. The researcher got $F_{\text{calc}}(4.16)$ was smaller than $F_{\text{table}}(4.24)$. Its mean $H_0$ was accepted and the nine classes was homogeneous. They completed some writing skills lesson and have advanced to the point where they should be able to compose a short essay. As a result, the authors sought out the most qualified students. And researcher choose 24 students in this VIII.1 class as participants. They completed some writing skills lesson and have advanced to the point where they should be able
to compose a short essay. As a result, the authors sought out the most qualified students.

Raters

As raters for the writing test, they utilized the services of two knowledgeable instructors from the same school. They were selected due to the fact that they shared comparable qualifications, as well as years of experience teaching and years of skill scoring high-stakes tests. Both of the raters had at least five years of experience working in the same school and each held a degree in English from a reputable university. In addition to that, they have instructed several students in English at the institution.

The researchers extended an invitation to the raters to take part in a training session that would last for two hours. The training consisted of an explanation of the rating method, a discussion of typical rating issues, and recommendations on how to minimize bias. The purpose of the training was to ultimately improve rating accuracy and rater agreement.

Rubrics for Evaluation

During the course of the inquiry, both a holistic and an analytical rubric were utilized. The holistic rubric is a scale with six points, and for each point, there is a broad explanation of the expected level of writing proficiency. The performance of students in their writing tasks is evaluated using a rubric that is modeled after the one that is used by the department instructors. In point of fact, the researchers devised it after conducting unstructured interviews with educators to get information about the criteria that are used by them when rating students’ written work. As a direct consequence of this, the suggested grading system incorporates two performance criteria, namely, topic comprehension and linguistic accuracy.

The analytical rubric, on the other hand, is a reworked version of Bachman and Palmer’s (1996). According to the findings of the researchers, the criterion-referenced rating scale developed by Bachman and Palmer to evaluate writing ability now includes a fifth sub-domain. This addition was inspired by issues that were context-specific. The end result is a scale with five points that consists of five different sub-domains of writing skill. These sub-domains are as follows: content, coherence, syntactic structures, vocabulary, and writing mechanics. There are a variety of clearly defined performance requirements that are specific to each area, and each rater is aware of these requirements.
Procedures for Rating

In the two sessions that immediately followed, each rater worked alone. Twenty (anonymous) writing examples were given to the raters for analysis during the first session, along with a copy of the holistic scoring rubric. Each essay received a single "holistic" score from the raters ranging from 0 to 5, with 5 representing the greatest score. Once it reached a total of 20, each participant's score was logged next to their participant number and increased by one point. The researchers received the graded writing samples and grading criteria within three days. In order to allow for a more impartial evaluation, the second session was planned to take place three days after the first. 20 identical writing samples that were all kept anonymous were distributed to each rater during this session, along with a copy of the analytic rubric. For each sub-domain of writing competence, they were instructed to provide a score on a range of 0 to 4 ranging from 0 (no knowledge) to 4 (complete knowledge), and then add all of these values together to arrive at a final score of 20. The researchers were given graded writing examples and grading criteria over the course of three days.

RESULT

To address the study's research questions, several statistical methods were employed. Using the holistic and analytic rubrics, the descriptive statistics of the students' scores were compiled initially. The descriptive statistics on each rater's evaluation of the writing sample using both rubrics were then provided. A t-test was conducted to determine if there was a significant difference between the means of the two scoring rubrics, and an analysis of variance was done to determine if the scoring judgments of the two raters for each of the two scoring rubrics differed substantially.

According to the findings, the standard deviation of the scores obtained using a holistic rubric was 4.162, while the mean score obtained using the holistic rubric was 14.69. The performance of the pupils was evaluated using the analytical rubric, which resulted in a mean score of 14.60 and a standard deviation of 3.983.

Table 1. Descriptive statistics for each of the scoring rubrics (N = 48)

<table>
<thead>
<tr>
<th>Rubrics</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>5</td>
<td>24</td>
<td>14.69</td>
<td>4.162</td>
</tr>
<tr>
<td>Analytical</td>
<td>7</td>
<td>24</td>
<td>14.60</td>
<td>3.983</td>
</tr>
</tbody>
</table>
The table below includes descriptive data for each of the two raters inside each of the two rubrics.

Table 2. Descriptive statistics for each of the two raters within each of the two rubrics (N = 24)

<table>
<thead>
<tr>
<th>Rubrics</th>
<th>Raters</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>Rater # 1</td>
<td>5</td>
<td>20</td>
<td>14.54</td>
<td>4.587</td>
</tr>
<tr>
<td></td>
<td>Rater # 2</td>
<td>7</td>
<td>20</td>
<td>14.83</td>
<td>3.784</td>
</tr>
<tr>
<td>Analytical</td>
<td>Rater # 1</td>
<td>6</td>
<td>20</td>
<td>14.50</td>
<td>4.128</td>
</tr>
<tr>
<td></td>
<td>Rater # 2</td>
<td>7</td>
<td>20</td>
<td>14.71</td>
<td>3.917</td>
</tr>
</tbody>
</table>

To evaluate whether or not there was a statistically significant difference between the means of the two grading rubrics, holistic and analytic, a t-test was carried out. The results of this test are presented in the following table. It was demonstrated that there was a difference that could be considered statistically significant between the two grading scales (t = 2.654, p < 0.05). According to the data, the analytical scoring rubric was discovered to be more difficult (M = 14.61, SD = 4.022) as compared to the holistic scoring technique (M = 14.68, SD = 4.185), which was determined to be more demanding.

It is essential that the score is not influenced in any way by the individual who is assigning the grade, and it is also essential that the results be comparable to one another. The dependability of the assessment is improved when the ratings that were given by the various raters may be compared to one another. In order to evaluate whether or not there were any significant discrepancies in the ratings that were supplied by the two raters, an analysis of variance was carried out on each of the several rubric techniques that were used. The findings demonstrated that there was not a significant difference between the two raters when applying an analytical rubric to grade the performances of the students; this was shown by the findings. When the raters used the holistic rubric, the results of their grading were drastically different from what they had been previously. In order to determine the reasons for the differences, a post-hoc analysis was carried out. According to the findings, there
was a difference between rater 1 and rater 2 at a significance level of P that was lower than 0.05.

It is of the utmost importance to conduct research into the relationship that exists between the two distinct scoring methods. When the correlation is strong, it implies that the two distinct approaches to scoring are likely to provide the same results. This is the case when the correlation is high. The value of 0.83 indicated that there was a highly significant link between the two, which led to the conclusion that there was a relationship between the two. On the other hand, the correlation in this kind of situation should be more than 0.90.

Table 3: Descriptive statistics for each pair of two raters, as well as the total score across the two scoring rubrics (N = 20).

<table>
<thead>
<tr>
<th>Pair 1 &amp; 2</th>
<th>Total Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>.23</td>
</tr>
<tr>
<td>Analytic</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>.21</td>
</tr>
</tbody>
</table>

**DISCUSSION**

The accomplishments of the students are currently being evaluated, and it is essential that this analysis be as precise as is humanly feasible because it may have ramifications for the young people being assessed (Black, 1998). Variation in any evaluation may be caused by a number of different variables; one of these is the raters' judgments of how well the pupils performed (Black, 1998). This was the principal focus of the investigation that was carried out.

They explored whether utilizing holistic or analytical scoring rubrics would be more beneficial for boosting the scoring reliability, and our findings indicated that there was a substantial difference between the means of the two scoring procedures, which were holistic and analytical. In particular, they discovered a substantial disparity between the means of the holistic scoring rubric and the analytical scoring rubric. In addition, the link between the two unique scoring techniques was examined and assessed. According to the findings, there was a link between the two criteria that may be regarded considerable. The findings of a subsequent study are addressed in further detail below; however, this does not mean
that the raters are in agreement. 0.80 was determined to be a sufficient correlation coefficient between the two unique scoring systems (Stemler, 2004).

Predict that the results of the students’ grading will be comparable regardless of who performs the grading. When evaluations from different raters may be compared to one another, the reliability of the evaluation is increased. The findings of an analysis of variance (ANOVA) demonstrated that there were no statistically significant differences between the three raters’ evaluations of the students’ achievements based on the analytical rubric. The students’ work was graded using the specified rubric. Nonetheless, when the raters used the holistic rubric, their scores were considerably distinct from one another, which indicated significant discrepancies between raters adopting holistic and analytical scoring rubrics. In order to study the link between holistic and analytical grading rubrics, Chi noticed these contradictions. After comparing the two various types of grading systems, Chi noticed these distinctions. Utilizing an analytical rubric resulted in more consistent and accurate results than did not using a rubric, according to the findings of this study.

Inter-rater reliability is distinct from intra-rater reliability, which refers to the dependability of an individual rater. Intra-rater reliability refers to the consistency of a single rater. Inter-rater reliability relates to the consistency of two or more raters. Intraclass correlation was utilized to evaluate both the consistency of the raters’ performance ratings and the inter-rater reliability of their evaluations. Given that the value of the Intra-Class Correlation Coefficient was more than 0.85, it is possible to conclude that the findings may be believed. According to Brown, Glasswell, and Harland (2004), the majority of research evaluating the reliability of intra-rater evaluations showed alpha values more than 0.70, which is regarded as usually excellent. In the majority of research that investigated the reliability of intra-rater evaluations, this was discovered. One study compared and contrasted the efficacy of holistic and analytical scoring rubrics. Analytic scoring provides students and teachers with quick and insightful feedback as well as particular information about the subject being evaluated test-taking abilities. Analytical scoring increases consistency between and among raters, according to numerous studies (Stoddart et al, 2000).
CONCLUSION

Multiple grading rubrics will be used to evaluate the student’s work on the different assignments they have been assigned. The effectiveness of students in respect to a predetermined set of requirements for a particular assignment is assessed using a grading strategy known as a rubric. Grades are given using this grading strategy. This article discusses the many advantages of using rubrics to assess student achievement on performance-based assignments. The following are a few of these advantages: By promoting learning and/or improving instruction and providing insightful feedback to students, using rubrics can have a positive impact on education. These benefits can be reached by: 1) improving judgment consistency when evaluating performance tasks; 2) offering a reliable performance evaluation judgment that cannot be obtained without the rubric; 3) having beneficial educational effects; and 4) giving students useful feedback. These advantages can be achieved by: 1) improving judgment consistency when evaluating performance tasks; 2) providing a valid performance assessment judgment that cannot be accomplished without the use of the rubric; and 3) decreasing the time needed to conduct performance assessments (Jonsson & Svingby, 2007; Archbald & Newmann, 1988).

This study evaluates the efficacy of two distinct types of rubrics, holistic and analytic, for measuring writing skills. Prior research examined the variations between widely-used rating systems for writing abilities. In addition, the need of incorporating rubrics for improved writing issue identification and score reliability was underlined. The current study concludes by evaluating the performance of two types of rubrics, holistic and analytic rubrics. This study aims to determine the type of rubric that is superior to others in terms of its capacity to assess a writer’s talents. The goal of this study is to examine the various methods in which analytic and holistic scoring rubrics are utilized to evaluate eighth-grade students' writing assignments.

The holistic and analytical scoring rubrics were contrasted in this study as trustworthy ways to grade writing for achievement objectives using various psychometric data. The results showed that, contrary to what was claimed in the literature, applying rubrics results in higher accurate scores. On the other side, it was discovered that analytical scoring provided even more reliable outcomes than comprehensive scoring schemes. Because the results can help teachers and students alike in recognizing students’ strengths and weaknesses as well as learning
requirements, analytical scoring seems to be highly beneficial in the classroom. As educators, we must acknowledge that the use of rubrics enhances the quality of the evaluation (Perlman, 2003).

The key contribution of this study is the tested hypotheses, which may be able to explain the variance in agreement between various grading models. Because of this, the same methodology may be used to test the same hypotheses with a larger and more representative sample of teachers, both in EFL and other subject areas.

Additionally, the study’s grading criteria only included written performance, thus a more diverse set of criteria, such as spoken performance, would have allowed for a more accurate comparison with the grading procedures used by teachers. A more varied material, however, could also be considered to emphasize the differences between the grading models by making the holistic grading even more complicated, perhaps leading to lower agreement.

As previously stated, the present study is unable to confirm the specific individual and contextual elements that affect teachers’ grading; it can only show that some variables outside of the grading model, such as allocating different weights to various criteria, cause variation in the sample. Given the wide variation in grades given, as well as the fact that this influence has consistently been found to be significant across a wide range of research over the years, the lack of guidance in this study is probably an artifact of the design. Since teachers are aware that their evaluations will be reviewed, it may be presumed that they will limit their written judgments to those they feel to be valid standards.

When considered collectively, it is advised that the results of this study be further studied using a bigger sample of teachers, more diversified content, and the inclusion of other subject areas. It is also advised to look into the possibility of supporting the reliability and/or validity of grading by training teachers to use strategies for evaluating and synthesizing evidence and addressing potential validity concerns.

REFERENCES


Putri & Melani / JELS 7 (2) (2022) 157-175


174


