International Journal of Language Education
Volume 4, Number 2, 2020, pp. 171-182
ISSN: 2548-8457 (Print) 2548-8465 (Online)
Doi: https://doi.org/10.26858/ijole.v4i2.13662

Exploring the Use of ESL Composition Profile for College Writing in the Indonesian Context

Lestari Setyowati
Faculty of Pedagogy and Psychology, Universitas PGRI Wiranegara, Pasuruan, Indonesia
Email: lestari.setyowati@yahoo.co.id

Sony Sukmawan
Faculty of Cultural Studies, Universitas Brawijaya, Malang, Indonesia
Email: sony_sukmawan@ub.ac.id

Ana Ahsana El-Sulukiyyah
Faculty of Pedagogy and Psychology, Universitas PGRI Wiranegara, Pasuruan, Indonesia
Email: aahsana3@gmail.com

Received: 03 May 2020
Reviewed: from 07 July 2020 to 09 September 2020
Accepted: 15 September 2020

Abstract
Assessing writing is a demanding task. If a lecturer of writing is not prepared with a reliable scoring rubric, the students' real performance might not be known. One of the well-known English as a second language (ESL) writing rubric is the Jacobs ESL Composition Profile which was developed by Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey in 1981, known as Jacobs ESL Composition Profile. This scoring rubric is popular among writing teachers and researchers to score students' composition. The present study is intended to 1) find out the internal consistency between raters who use the scoring rubric to assess the students' essay, and 2) describe the level of the students' writing performance when assessed by using Jacobs ESL scoring rubric, and 3) describe the raters' opinion when using the profile. The study uses descriptive quantitative design. The instruments to collect the data are documentation and interview. The data were collected in three months, from February to April 2020. The subjects of the study were two writing lecturers who taught English as a foreign language (EFL) and became raters for a research grant. The raters were asked to score 37 essays of the fourth-semester students. The result of the study shows that there is internal consistency between rater 1 and rater 2 when scoring the students' essay by using Jacobs ESL Composition Profile is high ($r. = 0.674$, $\alpha = 0.00 < 0.05$). The Cronbach alpha analysis also shows 0.722 which indicates a strong and high level of consistency. The students' writing performance fall in the average – good (moderate) level (49%), very good (high) level (40%), and only five students (11%) fall in the below-average – poor (low) level category. The result also reveals that Jacobs ESL Composition Profile is considered reliable to score essays even though it requires skills and practice because of its detailed description.

Keywords: ESL composition profile, reliability, rubric, writing,

Introduction

Scoring students' composition is a demanding task, especially in English as a foreign language context (EFL). Writing assessment is a complex activity that depends heavily on an individual judgment which is both time and energy consuming (González, Trejo, & Roux, 2017). EFL writing teachers are often caught in endless checks of students' errors in grammar, punctuation, spelling, diction, cohesive devices, transitions between paragraphs, main ideas, and sentence connectors. After checking all the students' writing, then, EFL lecturers are often undecided in terms of how the composition should be graded or which to be graded more. As stated by Yamanishi & Hijikata (2019), teachers often find difficulties not only in teaching writing but also in assessing the students' product. Thus, writing lecturers need to be equipped with sufficient knowledge with what to score and how to score the composition, so that he/she might not jump into a false conclusion about the students' writing performance.

This is why the scoring rubric plays an important role. A scoring rubric offers a clear mission of what and how to score the composition so that each writing element is graded consistently by the writing teacher/lecturer (Turgut & Kayaoğlu, 2015). As stated by Brookhart (2013), the main goal of the rubric is for performance assessment. She further states a rubric has two major elements, namely coherent sets of criteria and performance level descriptions for the criteria. Besides, as rubrics serve as indicators of learning outcomes, they help teachers/instructors to construct the observations and to match the observation with the descriptions to avoid a hasty judgment in a classroom situation (Bookhart, 2013).

In the second language writing, there are three types of scoring rubric commonly used by the writing teachers/ instructors to assess the students' composition, namely primary trait scoring rubric, holistic scoring rubric, and analytic scoring rubric (Weigle, 2002). Primary Trait Scoring Rubric (PTSG) focuses on specific rhetorical skills in writing (Latief, 1990; Weigle, 2002). Latief (2014:240) states that PTSG is simple and easy to use since it requires the rater to come up with only one single score to represent only the most important component of a text. He further states that this scoring rubric enables the researcher to train the raters rating the subjects' essay in a relatively short time. Yet, Setyowati (2016) states that PTSG is not adequately sensitive to differentiate well-developed composition and under-developed composition and might not be appropriate for scoring the students' final draft. Concerning this, Babbin & Horrison (1999) said that Primary Trait Scoring is mostly helpful in responding to the students' draft, and in encouraging and shaping revision.

The second type of scoring rubric is holistic scoring. The holistic scoring rubric describes the students' composition by applying all set of criteria at the same time and judge the quality of the composition in general (Martin-Kniep, 2000; Weigle, 2002; Bookhart, 2013). Holistic scoring requires the writing teachers/ lecturers to approach the composition in a holistic manner in which they give a single score for the whole text (Beyreli & Ari, 2009). Even though it is difficult to prepare, a holistic scoring rubric is easier and faster as it gives a single score for the whole composition (Beyreli & Ari, 2009) and suitable for summative assessment (Bookhart, 2013). Yet, holistic scoring has some controversies. First of all, a single score given to the text lacks information about what to improve in the composition and it is not suitable for the formative test (Bookhart, 2013). Secondly, if writing is taken as a whole, holistic scoring might not be able to give real information about the quality of the composition (Martin-Kniep, 2000). Some elements in the scoring may become more important than others. As a result, raters might reward a better score for not-so-good writing, and vice versa (Beyreli & Ari, 2009). Further, Martin-Kniep, (2000:35) states that a holistic scoring rubric cannot give much help to students' development particularly for those who have low and medium performance.

The third type of rubric is analytical scoring. The analytical scoring rubric describes each criterion separately and it works best for classroom purposes (Bookhart, 2013). The

criterion in the analytical scoring rubric should correspond with the elements of writing. Brown (2001) states that six major aspects of writing should be fulfilled by the writers in producing a written text. The six major aspects are content, organization, discourse, syntax, vocabulary, and mechanics. Content deals with the thesis statement, related ideas, development ideas, and the use of description. The organization element covers the effectiveness of the introduction, the sequences of ideas, and the appropriate length. The third element is a discourse that covers topic sentence, paragraph unity, transition, discourse maker, cohesion, theatrical convention, reference, fluency, economy, and variation. And the final element is mechanics which includes the use of spelling, punctuation, citation of reference, and appearance.

Perkins & Brutten (1990) state that writing ability should be learned as a whole rather than as a series as a separate element. According to Falk (1979), when a student writes a composition, he/she is involved with all aspects of language during the process of composing a piece of writing. He further states that there is a great chance that writing is learned holistically, instead of being learned separately through the mastery of separate skills. As Taylor (1981) points out, a dynamic interaction between content and language during the creative discovery of writing. Perkins & Brutten (1990) further say that the language aspects (content, organization, vocabulary, language use, and mechanics) interact with one another during writing.

One scoring rubric that accommodates the writing elements is Jacobs ESL Composition Profile. This profile was an analytical type and was developed by Jacobs and colleagues (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey) in 1981. In their rubrics, essays are rated on five different rating dimensions of writing quality with an a100-point scale, each having a different weight: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). Each set of criteria changes a four-level subjective judgment scale into interval scores. This profile is one of the most commonly used and dependable profiles for ESL composition rating (Lee et al, 2008) and its traits were designed by writing researchers working for a testing organization and is probably one of the most recognizable rubrics in the field of second language writing (Brooks, 2012). Furthermore, as stated by Ghanbari, Barati & Moinzadeh (2012), Jacobs ESL Composition profile has gained its popularity among second language teachers and researchers since its introduction in 1981.

Some research can be found in writing assessment, especially on Jacobs ESL composition Profile. The first research is conducted by Bacha (2001) who investigates the use of holistic and analytical scoring to grade the EFL students at Lebanese American University. All the students are Arab students who use English as a foreign language. The analytical scoring rubric used in the study was Jacob's ESL composition Profile. The result of the study shows that the analytical scoring rubric is a better choice to use for the EFL program as it gives more information about students' essay proficiency.

Secondly, Klimova (2011) compares the use of holistic scoring and analytical scoring developed by Jacobs *et al* (1981) in the Czech ESL context. The researcher concludes that analytical scoring is better than holistic scoring since it gives more information about the students' learning. Yet, she suggests that both types of scoring should be used when the teacher/rater is confronted with longer essays. Using both types of scoring can give more information about the overall progress and errors that can be avoided.

Thirdly, Ghanbari, Barati & Moinzadeh (2012) found out that Jacobs ESL Composition profile stands on the shaky ground in terms of how the five elements of the traits were originated. Based on their extensive research, their study reveals that the raters and commenters involved in the development of Jacobs ESL Composition Profile few of them came from a non-ESL background and had little experience in teaching English as a Second Language (TESOL). Meanwhile, the writers of the compositions were the freshmen at Cornell, Middlebury College, and the University of Pennsylvania, in which Ghanbari *et al* (2012) suspect, none of them were

second language learners. Ghanbari *et al* (2012) conclude that a scoring profile that comes from non-origin-ESL background might threaten its appropriateness when it is used to judge the EFL writing performance. They suggest that a local writing scale developed by local raters might lead to more valid outcomes because it is context-based.

Yet, previous research on Jacobs ESL Composition Profile has not given enough information about the internal consistency between raters and their perception when using it in the EFL context. Investigating the internal consistency between the raters can give valuable information to the writing teachers and raters about the validity of the instrument to score the writing and whether it is indeed reliable and does not cause misinterpretation when using it. The result of the study is also expected to give insights in terms of what to do to improve the internal consistency between raters so that the real performance of the students can be acknowledged. Therefore, the present study is intended to 1) find out the internal consistency between raters who use the scoring rubric to assess the students' essay, 2) describe the level of the students' writing performance when assessed by using Jacobs ESL scoring rubric, and 3) describe the raters' opinion when using the profile.

Research methods

Research design

This research uses a mixed-method approach because it employs both the qualitative data and quantitative data (Creswell, 2013). The mixed-method approach allows the researchers to integrate two forms of data to have a better understanding of the research problem rather than using only one approach (Creswell, 2013). The participants of the research were two writing lecturers and the 37 fourth semester students of the English Education study program, Faculty of Pedagogy and Psychology, Universitas PGRI Wiranegara, Pasuruan, East Java, Indonesia.

The raters are the writing lecturers in the English Education Study program who have five years of experience in teaching writing and developing writing tasks. Both raters are female and have bachelor degrees in English language education and master degrees in English education. They have seven years' experience of teaching English for college-level in the institution. Both of them work in the English Education study program of the University PGRI Wiranegara. One rater had received a young lecturer research grant from the Indonesian government in 2019 with a project focusing on essay writing for EFL learners. Meanwhile, the other rater once became a research member of a young research grant from the Indonesian government in 2018 focusing on sentence-level writing for EFL learners.

Data collection

The instruments used were interview and test documentation. The data was collected from February – April 2020. The students and the raters were told that they would be involved in a research project. The students' essays were collected after they were assigned to write an essay. The researchers gave the raters two weeks of training before the actual writing task. Training the raters to score similar composition by using a similar rubric could improve the accuracy of assessment and lessen the differences. The training mainly emphasized raters' understanding of the intended rating criteria. In the training session, some activities were elaborated as follows. The purpose of the test was explained to the two raters. Then each point of the scoring criteria was explained provided with an example. During the training, the raters were assigned to practice rating some compositions samples of the same person to be rated. The raters were given two weeks to finish scoring the students' papers. This length of time was given to avoid the errors of either overestimating or underestimating the true level of students' skill being assessed that are coming from the raters' physical and emotional constraints. Two

weeks was considered long enough as usually in the semester final examination every lecturer has two weeks to finish scoring all students' papers. The set of scores obtained from Jacobs ESL Composition Profile has ten points tolerated discrepancy. This meant that if there were a discrepancy of more than ten points, a third rater was invited to score the same essay. The scores used for computation are the closest score between raters with ten range points.

Data analysis

The quantitative data analysis was used to find out the internal consistency between raters and the students' level of proficiency. The researchers also used descriptive statistics to present the numerical data in graphs and percentages (Ramsey, 2011). On the other hand, the qualitative analysis was used to analyze the data gained from the interview and chat messages. Since some of the data are qualitative data, the researchers employ simple codification to categorize the qualitative data

The students' writing is scored by using Jacobs ESL composition profile (Jacobs, et al, 1981). The rubric has five different rating categories of writing quality with an a100-point scale. They are content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). Each element has a different way of scoring with the description attached to each category. After all the students' composition is graded by the raters, the scores are then grouped. The researchers employ the university standard to find out the students' level of writing performance. The university standard is used to grade the students' performance for all subjects. The criteria range from excellent (91 – 100), very good (84 – 90), good (77 – 83), average (71 – 76), below average (66 – 70), poor (61 – 65), very poor (55 – 60), and fail (45 - 54). The researchers, however, categorize the students' performance into three main levels, namely high (very good-excellent), moderate average-good), and low (poor-below average). The high-performance criteria fall in the category 84-100, the moderate category falls in 71-83, and the low category falls in 61-70. Two measures of inter-rater reliability are used in this study, namely Pearson Product Moment Correlation between the first and second-raters, and Coefficient Alpha which provides an estimate of the internal consistency of the final scores based upon two raters per essay. The correlation of the scores of the two raters and the coefficient alpha was computed by a means of SPSS IBM 20 for Windows.

Finding and discussion

Internal consistency between raters

The students' compositions were rated by two raters independently using Jacobs ESL Composition Profile (Jacobs *et al.* , 1981) using five criteria, namely content, organization, vocabulary, language use, and mechanic. Table 1 describes the descriptive statistics of the scores obtained from the raters.

Table 1. Descriptive Statistics

	N	Min	Max	Mean	Std. Deviation
R1	37	63.00	94.00	81.1622	9.39714
R2	37	68.00	88.00	80.6216	5.10093
Valid N (listwise)	37				

Before the computation of descriptive statistics, the raw data shows there was nine students' essay, out of 37, which were either overrated or underrated by both raters. Since the discrepancy between raters was more than ten points as agreed in the beginning, a third rater was invited to score nine students' essays. The closest scores among the raters were used for the computational analysis. According to Bacha (2001), if there is a score discrepancy, a third rater can be invited to score the same composition, and the closest score is used to get the average score. Table 1 shows the mean scores obtained from Rater 1 is 81.1622, and the mean score from rater 2 is 80.6216. It indicates that the mean score obtained from Rater 1 is higher 0.5446 point than the mean score from Rater 2. In terms of the standard deviation, the standard deviation from rater 2 (5.10093) is considered smaller than the standard deviation of the scores from rater 1 (9.39714) and has a 4.29621 range in standard deviation. Even though the mean of the raters differs only 0.5446 point range discrepancy, the standard deviation shows that the scores given by the rater 1 have more variation and spread farther away from the mean, unlike rater 2.

Table 2. Correlations

		R1	R2
R1	Pearson Correlation	1	.674**
	Sig. (2-tailed)		.000
	N	37	37
R2	Pearson Correlation	.674**	1
	Sig. (2-tailed)	.000	
	N	37	37

**. Correlation is significant at the 0.01 level (2-tailed).

The computational analysis of the correlation coefficient between rater 1 and rater 2 shows the degree of strength. Table 2 shows that the reliability coefficient is 0.674 which indicates a high and strong level of consistency between the first and second-raters and significant at 0.05 (α. 0.00 < 0.05).

Table 3. Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.722	.805	2

The reliability analysis by using Cronbach alpha also shows similar results in terms of internal consistency. The computational analysis of Cronbach Alpha value results in .722 which shows high consistency and reliability between rater 1 and rater 2. According to Sujarweni (2014), Cronbach alpha value which is higher than .60 is considered reliable and consistent.

In reality, although raters are trained with the scoring rubric to assess writing performance, all raters can be a little accurate when doing it. According to Wang (2009) at least there are two ways to improve inter-rater reliability. The first one is raters training, and the second one is minimizing subjectivity during the rating process. Wang (2009) and Zhang, Xiao, & Luo (2015) state that high inter-rater reliability between scorers/raters can be gained through training. They believe that training the raters to score similar composition by using a similar rubric could maximize the accuracy of writing assessment and minimize the differences as a result of different backgrounds.

The raters involved in the study have undergone two weeks of training to use Jacobs ESL Composition Profile before the actual rating. Yet, it seems that two weeks of training is

not enough and more time needs to be devoted to training to make the raters become more and more familiar with each criterion and become accustomed to it. Furthermore, according to Zhang, Xiao, & Luo (2015) to achieve high reliability, the raters need to have adequate knowledge and skill in the language testing area. The raters involved in the present study are the writing teachers with limited knowledge of language assessment. This might be one of cause why the initial scores produced by the raters are a little bit inconsistent, which resulted in the invitation of the third rater to rescore the same composition.

The second strategy to improve inter-rater reliability is to minimize subjectivity during scoring by identifying subjects by numbers, instead of name (Wang, 2009). Identifying subjects by name would affect the way the raters to score. For example, the rater may be influenced by gender or subjects' familiarity. If this happens, the raters inevitably had expectations for particular subjects which later cause subjective scoring. Thus, identifying subjects by number would reduce such effects. In the present study, the subjectivity case might happen as both raters are familiar with each of the student's names. Therefore, before the actual rating, each of the students' names should be covered with a correction pen, or better yet, the students are ordered to write their attendance list number only, and not to give their name on their writing sheet.

Students' level of essay writing performance

The students' level of writing performance is gained through the average score between raters. The spread of the students' score is presented in Graph 1

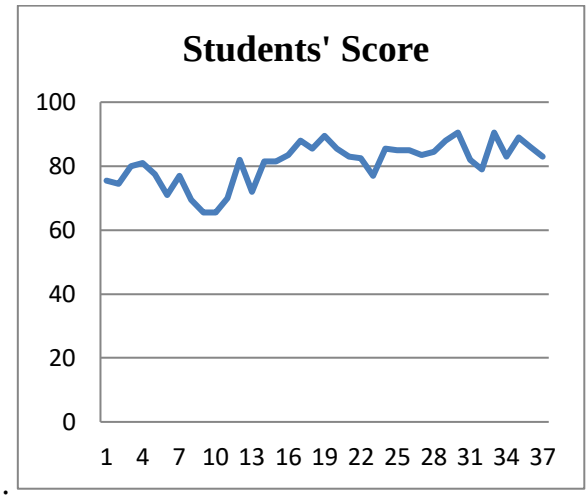

Figure 1. Students' score

Figure 2 shows that none of the students fall below 60, and none reaches 91 as the minimum level of excellent criterion. As stated above, the students' are grouped to level based on the institutional standard. The students who get 84 above (A- to A) belong to high (very good – excellent) level writing performance. Those who get 71-83 (B to B+) go to moderate (average – good) level writing performance, and those who get 61-70 (C+ to B-) belong to low (poor – below average) level writing performance. The students' level of essay writing performance is presented in graph 3.

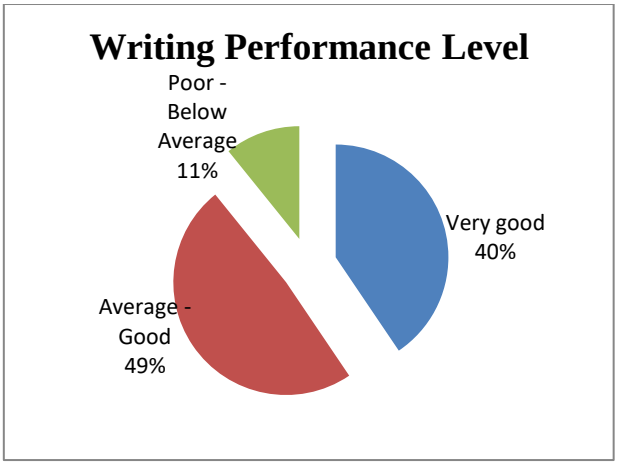

Figure 2. Students' level writing performance.

The data shows that the moderate level (49%) and the very good level (40%) writing performance when rated by Jacobs ESL composition profile is almost equal. Unfortunately, there are still 4 persons (11%) belongs to the low-level writing performance.

The data shows that some students still need some help to improve their writing performance. To achieve this, the writing teacher/lecturer can firstly respond to the students' drafts by using PTSG before it is finally assessed by Jacobs ESL Composition Profile. As stated by Babbin & Harrison (1999 Primary Trait Scoring is more helpful in assessing the writing draft, and in encouraging revision. Furthermore, Klimova (2011) also suggests that in assessing longer essay, the teacher/rater need to utilize two types of scoring if necessary.

Raters' opinion in using Jacobs ESL composition profile

Before being introduced to Jacobs ESL Composition Profile, the writing lecturers who became the raters for the research grant rated the students' scores mostly based on judgment. The use of judgment in scoring compositions will result in unreliable performance assessment of the students' work (Gonzales, Trejo, Roux, 2017). They suggest that any assessment should be free from bias and reliability.

> I used to use feelings to score the students' writing. I had no particular scoring rubric. The first thing I used to see was their grammar and how the sentences connect. (AA/Rater1)

> I used to make my rubric, but my rubric is feeling based. I assessed the students' composition based on three elements, content (40%), grammar (30%), and vocabulary (30%). (DA/ Rater 2)

After the raters were introduced to Jacobs ESL Composition Profile they could focus the assessment on what to score and how to score the composition based on each criterion. It was not surprising that they have some problems in the beginning in terms of how to use the rubric. They confessed that the problems laid in the detailed elements to score.

> Jacob's ESL Composition profile has too detailed descriptions. Teachers who were first introduced with the rubric might have some difficulties in the beginning. But if it is used often, the teacher will get used to it. I think he/she needs only to see the subscore range in each criterion. (DA/ Rater 2)

> It is difficult to use Jacobs in the beginning. I have difficulties matching my judgment with the subscores of each element in the rubric For example, in the beginning, it's difficult to

> differentiate the language use score and the vocabulary score. I think in the content criteria, I'm too stingy to give the score. (A/Rater 1)

Jacobs ESL Composition Profile indeed details in its descriptions for each element. That might cause some problems for those who use the rubric for the first time. Yet, if the raters are given enough practice, they will be proficient. As discussed earlier, having some training in using a particular scoring rubric to judge the students' writing may result in better reliability (Wang, 2009; Zhang, Xiao, & Luo, 2015). The raters also confessed that Jacob's ESL Composition Profile helped them in better judgment.

> "Now, I get used to using Jacobs ESL Composition Profile. It helps me to score the students' writing, and help me to become more objective. Overall, Jacobs is very useful to measure the students' writing products" (AA/Rater 1)

> Overall, I think Jacobs's ESL Composition Profile is good to use to score the students' writing. (DA/ Rater 2)

The result of this research is similar to Gonzales, Trejo, Roux (2017). In their research, the raters in the study show a positive attitude toward the use of a scoring rubric. The majority of the respondents in their research report say that the scoring rubric makes the assessment process more efficient and more practical. They further conclude that teachers should use rubrics to score the students' writing. Yet, the use of rubric is not a guarantee in objectivity. This is so unfortunate because when assessing writing, the quality of the students' writing is influenced by the raters' judgments. Gonzales, Trejo, Roux (2017) argue that writing assessment depends not only on the raters' subjectivity but also on their interpretation and the rubric in use.

The purposes of grading practice, among others, are to provide feedback, to motivate the students, to compare their performance, and to measure learning (Schinske & Tanner, 2014). Because of these, raters must score the students' writing objectively. Yet, when it comes to scoring writing, it is hard not to be subjective. As stated by Attali (2015), grading subjective assessment, like writing, raters may give different scores when scoring the same item and it may vary drastically depending on their experience, length of training, and bias. Thus, raters need a scoring rubric to enable them to have the same voice in judging the students' writing performance.

The need to reduce the raters' inaccuracy in scoring is necessary to make the data valid. Kim (2015) states that there are three ways to increase the scoring reliability in a subjective assessment, namely giving raters some training to use the particular rubric, selecting raters who are experienced in grading a test, and increasing reliability. Similarly, as Boyer (2020) points out, there are three ways to mitigate the impact of rater's inaccuracies. She divides the mitigation into three phases, before, during, and after scoring. She suggests that before scoring, the scoring rubric should have detail scoring criteria. The rubrics should explicitly describe the expected responses from the students. If the rubric uses unclear language, raters might have difficulties in their judgment and in making distinctions to decide the level of the performance. The second phase in mitigating rater inaccuracy is providing training and monitoring the process of scoring. One common approach is to show the model of scoring of a particular response based on the intended rubric. And the last phase as suggested by Boyer (2020) is using the statistical procedure to detect and correct the inaccuracy. Taken into this context, Jacobs ESL Composition Profile has fulfilled the criteria of being a detail and specific scoring traits in which each element is clearly described. Thus, the chance of raters inaccuracy can be minimized.

In sum, the researchers suggest that applying a detailed and reliable scoring rubric like Jacobs's ESL composition profile alone is not enough. Since scoring writing is mostly done by human instruments, the issue of subjectivity is always present. The user of the rubric needs to be experienced enough in making the best use of it. The students' writing performance needs to be fairly evaluated and graded. The false judgment of their performance can influence their academic life in the future. To avoid the false and unfair scoring, the teachers, or any writing instructors, should enrich themselves with assessment literacy, or joining training and professionalization concerning writing.

Conclusion

The result of the study allowed the researchers to conclude that Jacobs ESL Composition Profile is still considered a reliable scoring rubric to be used in EFL settings despite its controversies. Even though Jacobs ESL Composition ESL Composition Profile is almost two decades old in its development, it is still proven to be a reliable tool to score writing. To improve the reliability of the scoring, the raters can take some training in using the rubric, as well as improve their knowledge on language testing and writing assessment course. Giving them opportunities for teachers/lecturers with such training can improve the raters' reliability so that the students' real performance can be identified. Since this study is a small scale in nature, the result of this study cannot be used as a generalization. All in all, using a reliable scoring rubric can give a fair game for both the teachers and the students.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

Funding acknowledgment

This research is fully supported by the Indonesian Directorate of Research and Community Engagement (DRPM) of Higher Education for year funding in 2020

References

Attali, Y (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. Language Testing 33, no. 1, 99–115. doi:10.1177/0265532215582283

Babbin, E. H & Harrison, K. (1999). *Contemporary Composition Studies: A Guide to Theorists and Terms.* London: Greenwood Publishing Group

Bacha, N. (2001). Writing evaluation: What Can Analytic Versus Holistic Essay Scoring Tell Us? *System*, 29, 371- 383.

Beyreli, L & Ari, G. (2009). The Use of Analytic Rubric in the Assessment of Writing Performance-Inter-Rater Concordance Study. *Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice*. 1,105-125. Retrieved from https://files.eric.ed.gov/fulltext/EJ837777.pdf

Brooks, Gavin. (2012). Assessment and Academic Writing: A Look at the Use of Rubrics in the Second Language Writing Classroom. *Kwansei Gakuin University Humanities Review*. 17, 227-240. Retrieved from http://kgur.kwansei.ac.jp/dspace/bitstream/10236/10548/1/17-18.PDF

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading.* Alexandria: Association for Supervision & Curriculum Development

Brown, H. D. (2001). Teaching by Principles: An Interview Approach to Language Pedagogy, Second Edition.New York: Longman, Inc.

Boyer, M. (2020). *Understanding and Mitigating Rater Inaccuracies in Educational Assessment Scoring. Center for Assessment*. Retrieved from https://www.nciea.org/blog/educational-assessment/understanding-and-mitigating-rater-inaccuracies-educational-assessment

Creswell, J. W. (2013). Research design: Qualitative, quantitative, and mixed methods approaches. Thousand Oaks, CA: Sage Publications

Falk, J. S. (1979). Language Acquisition and the Teaching and Learning of Writing." *College English* 41, 436-47.

Ghanbari,B., Barati, H., & Moinzadeh, A. (2012). Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny. *Language Testing in Asia.* 2 (1), 83-100

González, E. F., Trejo, N. P. &Roux, R. (2017). Assessing EFLuniversity students' writing: a study of score reliability. *Revista Electrónica de Investigación Educativa,* 19(2), 91-103.https://doi.org/10.24320/redie.2017.19.2.928

Jacobs, H. L., S. A. Zingraf, D. R. Wormuth, V. F. Hartfiel, and J. B. Hughey. (1981). *Testing ESL Composition: A Practical Approach.* Rowley, MA: Newbury House.

Kim, Hyun Jung (2015). A Qualitative Analysis Of Rater Behavior On An L2 Speaking Assessment. *Language Assessment Quarterly*. 12 (3), 239–61.

Klimova, B. F. (2011). Evaluating writing in English as a second language. *Procedia - Social and Behavioral Sciences* 28, 390 – 394

Latief, M.A. (1990). *Assesment of English Writing Skills for Students of English as aForeign Language at the Institute of Teachers Training and Education IKIP Malang Indonesia*. Unpublished Ph.D. Disertation. Iowa: Graduate College of the University of Iowa.

Latief, M.A. (2014). *Research Methods on Language Learning: An Introduction* (2nd Ed). Malang: Universitas Negeri Malang (UM Press)

Lee, Y W, Gentile, C, & and Kantor, R. (2008). *Analytic Scoring of TOEFL® CBT Essays: Scores From Humans and E-rater.* Princenton: ETS

Martin-Kniep, G. O. (2000). *Becoming a Beter Teacher: Eight Innovations That Work.* Alexandria: Association for Supervision & Curriculum Development.

Perkins, K & Brutten, S. R. . (1990). Writing: A Holistic Or Atomistic Entity? *Journal of Basic Writing* . 9 (1), 75-84

Rumsey, D. (2011). *Statistics Workbook For Dummies, Statistics II For Dummies,* and *Probability For Dummies*. New Jersey: John Wiley & Sons

Schinske, J.,& Tanner, K 2014. Teaching More by Grading Less (or Differently). *CBE Life Sci Educ.* 13(2): 159–166. doi: 10.1187/cbe.CBE-14-03-0054

Setyowati, L. (2016). *The Effect of Planning on EFL Students' Writing Performance Across Different Levels of Self-Efficacy.* Unpublished Ph.D. Dissertation. Malang: State University of Malang.

Sujarweni, V.W. (2014). *SPSS Untuk Penelitian.* Yogyakarta: Pustaka Baru Press

Taylor, B. P. (1981). Content and Written Form: A Two-Way Street." *TESOL Quarterly* 15, 5-13.

Turgut, F., & Kayaoğlu, M. N. (2015). Using rubrics as an instructional tool in EFL writing courses. *Journal of Language and Linguistic Studies,* 11(1), 47-58.

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

Wang, P. (2009). The Inter-rater Reliability in Scoring Composition. *English Language Teaching,* 2(3), 39-43. Retrieved from www.ccsenet.org/journal/index.php/elt

Yamanishi, H., Ono, M. & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment. *Lang Test Asia* **9,** 13 https://doi.org/10.1186/s40468-019-0087-6

Zhang, B. Xiao, Y., & Luo, J. (2015). Rater Reliability and Score Discrepancy Under Holistic and Analytic Scoring of Second Language Writing. *Language Testing in Asia*, 5 (5). https://doi.org/10.1186/s40468-015-0014-4