Students' writing test: an argumentative study of English as a foreign language learner

Masrul Masrul[1], Santi Erliana[2]

[1]Department of English Language Education, Faculty of Teacher Training and Education, Universitas Pahlawan Tuanku Tambusai, Bangkinang, Indonesia

[2]Department of Tadris in English, Faculty of Tarbiyah and Teacher Training, Institut Agama Islam Negeri Palangka Raya, Palangka Raya, Indonesia

Article Info

Article history:

Received May 24, 2023
Revised Nov 2, 2023
Accepted Dec 22, 2023

Keywords:

Argumentative writing
Assessment writing
Language acquisition
Writing proficiency
Writing test

ABSTRACT

Writing is hard for students who are learning English; they often find it challenging to transform what is on their mind in writing. Therefore, this study examined the relationship between the writing test and assessment writing through argumentative writing. Data was analyzed using the correlation test to determine the close relationship between independent and dependent variables. This study involved 100 students from the Department of English Education at the University of Riau, Indonesia. The results showed that the writing test and assessment writing was closely related, as evidenced by the influence and significance between the writing test and assessment writing, which was tested through argumentative writing. The results revealed that the writing test and assessment writing have similar results. Overall, both variables are equally important and related.

Corresponding Author:

Masrul Masrul
Department of English Language Education, Faculty of Teacher Training and Education
Universitas Pahlawan Tuanku Tambusai
Tuanku Tambusai Street No. 23, Bangkinang, Kampar, Riau, Indonesia
Email: masrulm25@gmail.com

1. INTRODUCTION

Writing is difficult for students and teachers. One of the most common methods for determining an English language learning (ELL) student's competence and knowledge level is to have them compose an essay or present a writing sample. Evaluation holds a crucial role in the university student journey as the initial course placements profoundly influence the chosen academic paths, duration of enrollment, and, ultimately, the probability of attaining educational objectives. As per research, English as a second language (ESL) placement significantly impacts ELL students' access to and success in higher education [1]. A previous study found that enrolling students in community college ESL classes can delay their enrollment in college-level coursework and hinder their success in transfer-level courses, delaying their ability to achieve an associate's degree or other credential [2].

Many standardized writing assessments include essay writing as a constructed-response problem. The spontaneous, scheduled nature of the essay writing examinations has triggered increasing criticism for their lack of realism compared to real-world writing in classrooms and workplaces [3]. The growing use of automated essay scores to grade these items has raised additional concern about essay writing evaluations failing to accurately reflect the writing form [4]. Teachers of writing have raised concern that the absence of a diversity of genres on these timed exams, and the lack of opportunity for student learning in the writing processes, is having a negative impact on writing instruction.

The scoring methods' reliance on the identification of superficial faults in writing classes, as well as their proclivity for producing useless techniques, has been blamed for these difficulties. With the growing criticism and concerns over the construct-based interpretations and consequences of timed essay exam, more data is needed to support the accuracy of higher-grade exams using essay type collaborative reasoning (CR) questions. Moss [5] has emphasized that assessments differ in the range of educational goals they support and document. We must analyze the "consequential validity" of standardized exams before assigning them high-stakes. The greater social ramifications of utilizing a particular test for a certain purpose are the consequential element of validity. As a result, the current study examined the relationship between the writing test and assessment writing through argumentative writing.

Investigating relationships between standardized tests and outer requirements of involvement, such as total score on some other tests (same or distinct fabrication), curriculum sub-scores, cumulative test scores, and comprehensive conviction of learning achievement, and thus attempting to address the various strands of construct, is a popular form of authenticity data proving an evaluation use assertion [6]. External variables, such as learners' self, instructor evaluations, educational classification, and specific institutional tests, were all used to moderate the relationship between testing results and student achievement indicators [7]. Several more conversational analysis theories argue for an undergraduate perspective of assessments, recognizing ideas like truthfulness, exam subject matter, and instructional significance as demonstrating the factors that should be discussed in a single, overarching rationalization for construct validity. This viewpoint has been demonstrated to be as important for checking as it has been for other domains of evaluation [8].

Plakans [9] discovered that writing-only and read-to-write tasks elicited different composition behaviors on either side, stimulating a more active approach. The formulation of test tasks has also been influenced by studies that used student and instructor opinions to determine the significance and utility of test activities for academic success [10]. The goal of the test was to determine how well pupils could communicate in English for academic purposes. Once, trying to compare crucial grammatical structures, discursive characteristics, and vocabulary utilization between the study variables, an in-class speaking task, and an out-of-class speaking task, the researchers discovered a partial overlap, prompting some rebuttals to the assessment's overall validity claim.

Brooks and Swain [11] examined spoken information from a large English as foreign language (EFL) exam. The test aimed to see how effectively students could speak in English lessons. The researchers identified a significant agreement by examining major grammatical characteristics, conversational characteristics, and word use throughout the study variables, the discussing task, and the talking task, leading to some rejoinders to the assessment's general accuracy claim. However, Weigle and Friginal [12] compared linguistic aspects of spontaneous test essays to the characteristics of good students' writing on academic projects in various disciplines and multiple fields using a multidimensional analysis technique. For some fields, such as arts and humanities, there was more overlap between the properties of the two corpora than for others, such as health and natural sciences. Researchers examined the examination writings prepared concerning various test of EFL (TOEFL) individual questions and discovered that the overlapping between both the various corpora's features was largely affected by the issue.

Typically, all applicants to graduate school must take the graduate record examination (GRE), and those whose first language is not English must also take the TOEFL test. Both evaluations include literature as one of the elements, and both include two types of exercises lasting 30–45 minutes. The activities that better reflected the support-a-position-on-an-issue (argumentative) genre in this study were the GRE issue and TOEFL independent exams. The topic pressure from various examinees to assess a subject and generate arguments using explanations and evidence to defend the author's viewpoint, while the TOEFL individual activity requires students to advocate a viewpoint on a subject [13].

A genre-based approach to teaching writing assumes that preparing students for specific writing tasks necessitates explicit instruction in the key characteristics that distinguish these types of writing from one another [14]. Differences between genres include characteristics that identify the genre's social and communication environment. The rhetorical profiles differ by writing genre, and this research looked at the logical characteristics associated with argumentative writing, an instructional crucial type to include in elevated exams [15]. To support the validity argument for timed essay writing tests and strengthen the assessment use argument for the writing assessments studied, researchers conducted a study examining the relationship between the writing test and assessment writing through argumentative writing.

2. METHOD

In this study, data were analyzed using the correlation test, one of the statistical tests used to determine the close relationship between independent and dependent variables. In the correlation test, the determination of the strength or weakness of a relationship is assessed based on the value being closer to 1 or -1. However, closer to 0 means the relationship between the two variables is weak. In the correlation test, several tests can be used, namely Pearson, Kendal's, and Spearman's. In this study, the Pearson correlation test was used.

After the correlation test had been carried out, the Chi-square test was conducted to examine the relationship between the two calculated variables (count data). The basis of the test used is the difference in the proportion value of the observed value and the expected value. Some associate the Chi-square test as a test to see the relationship between two qualitative (categorical) variables. Generally, the relationship between two qualitative variables is descriptively displayed as a contingency table (cross tabulation). There are many types of proportion difference test/Chi-square tests proposed by many authors and literature, each type of test is based on certain assumptions that must be met by the data to be tested. Pearson Chi-square test is used to test the relationship between two categorical variables where the assumption is that the expected value for each cell is at least 5 or more, in other words the more data needed for the Pearson Chi-square test.

Most of the data for this study came from two groups of participants. The total number of participants was 100 from the Department of English Education, University of Riau. This research was conducted on first-semester students in the Department of English Education. Examination impressions of the state of the written test were collected through a questionnaire administered to candidates who had recently completed the writing test.

The teachers were the last group to provide information, and they took part in focus groups and completed surveys. The workings and consequences of the writing test were examined in the focus group interviews. After the focus group discussion, the instructor filled out an anonymous questionnaire after the interview to learn more about how the writing test was administered, how assessment writing was conducted, and how it impacted their writing results. Although these discussions were not recorded, stakeholders were provided with notes and summaries at a later date to ensure that they accurately reflected their experiences and opinions.

The study's data came from student writing under three conditions and situations: test writing (assessment writing), course-related writing samples from various genres (instructional writing), argumentative writing samples from several academic subjects (argument writing across fields of study), and opinion articles from various writers [16]. In the process of assessing written data, the researchers employed essays that received high scores, specifically those with scores of 4, 5, and 6 on either a 5-point or 6-point scale. Writing samples in the genre of argumentation, developed as part of an advanced writing course focused on teaching students how to write successfully in multiple genres, were the key academic external criterion. Researchers used grade-A writing samples in the argument genre performed for course assignments in various academic disciplines by students in the Department of English Education for the second academic external criterion.

Researchers used essays from two different large-scale high-stakes examinations that were composed in response to two different tasks. Both writing assessments were created to evaluate the students' abilities to write. The GRE essays were written in response to the issue task, which requires the test taker to evaluate an issue and develop arguments that support the writer's point of view on the subject using reasons and examples. Reduced essays fail to acknowledge and perform accordingly, commonly due to defects in the text component of the process or misinterpreting the topic/task. Scores 4–6 corresponded to expository writing that effectively handled the task of attempting to express a personal viewpoint on a problem supported by arguments; low-scoring essays fail to acknowledge the task appropriately, frequently due to flaws in the text component of the process or misinterpretation of the topic/task.

The pool was limited to submissions that received the same score from two independent raters for each score level (4, 5, and 6), and one essay per prompt was chosen randomly (112 entries for each score level). Similarly, for the independent assignment, in which test takers must support their opinion on a topic, researchers chose two sets of TOEFL essays, one for each score level of 4 and 5. This task is scored on a 5-point scale. For various reasons, argumentative writing has been chosen as the focus. It is necessary for academic performance as well as everyday life. Argumentative writing has a reputation for being difficult [17]. Recent interpretations, however, call into question both this viewpoint and the concomitant belief that young writers should not be assigned argumentative writing. Given the importance of argumentative writing, various perspectives on its difficulty, and competing perspectives on how to teach writing, it seems worthwhile to investigate the roles of development, direct instruction, and experience in the development of expertise in this type of writing [18].

Researchers sampled a single item per topic from the New York Times topic index, keeping the articles under 2,000 and 200 words. Researchers maintained track of the authors and resampled from the provided topic if an article from the same author on a different topic had previously been chosen. A total of 920 articles were produced as a consequence of this approach. Researchers believed this corpus covered a wide range of public-interest themes without focusing on one in particular. Argumentative (i.e., supporting a position on an issue) writing is crucial for everyone.

First-semester outcomes for both listening-speaking and reading-writing students were utilized to assess the writing sample's capacity to predict student performance. The utilization of letter grades or grade point averages (GPAs) in these inquiries has been recognized as problematic because of the limited scope of outcomes. Instead, the results have been expressed as a final percentage. Additionally, where instructors provided appropriate content, non-target language use (TLU) domain competencies had limited impact on final outcomes. As an illustration, evaluating factors such as attendance, participation, and imposing penalties for late submissions do not accurately represent language proficiency, an aspect that could be addressed by a writing test. As a result, those items' influence on course results was reduced and disclosed where possible [19].

The assessment that examinees who earn higher ratings on the assessed skills demonstrate a greater understanding of the construct under consideration than those who receive lower ratings [20]. Furthermore, the outfit mean square estimations for each rating category are determined to be within the allowed range, implying that the rating scales are working well [21]. While average candidate ability rises with each step up the rating system, there is reason to be concerned about one or both of the following: this indicates that examinees who receive higher ratings on the measured skills demonstrate more of the construct being examined than those who receive lower scores [20].

Furthermore, the outfit mean square estimations for each rating category are determined to be within the allowed range, implying that the rating scales are working well [22]. While average candidate ability rises with each of the rating system, there is reason to be concerned about one or both of the following: This indicates that examinees who receive higher ratings on the measured skills demonstrate more of the construct being examined than those who receive lower scores [20]. Furthermore, the outfit mean square estimations for each rating category are determined to be within the allowed range, implying that the rating scales are working well [23].

While average candidate ability rises with each step up the rating system, there is reason to be concerned about one or both of the following: the quantity of discernible proficiency levels identified among test takers does not align with the quantity of scale levels; the implemented writing test at the institution is capable of consistently distinguishing between more scale levels than the strata of proficiency levels, and the raters may not uniformly comprehend or consistently apply the rating scales [24]. At least two English professors graded each composition. The scoring was carried out using a locally devised analytic scoring rubric. Support, organization, sentence variability, diction, and grammar errors were among the requirements [25]. The institution has no additional information or documents detailing the criteria. To assist the rater in judging application writing, several descriptors were provided for each criterion, and the rater checked the description that best described the candidate's work [26]. There were six descriptors available for each criterion, aligning with scores ranging from one to six. The final score of 6 was obtained by averaging the results of all criteria.

3. RESULTS AND DISCUSSION

3.1. Descriptive statistics

Descriptive statistics serve as an initial method of analyzing data, offering a summary of the measured variables. In descriptive statistics, the analysis can manifest through measures of data central tendency (mean, mode, and median) and data dispersion (standard deviation, variance). The mean and standard deviation of all variables in the study are presented in Table 1.

Table 1. Descriptive statistics of research variables

Variable	Mean	SD	MnSq	Z Std
Argument	3.62	1.53	15.35	6.12
Opinion/position	3.75	1.56	16.41	6.31
Issue	3.62	1.67	15.79	6.35
Topic	3.49	1.54	14.52	6.02
Reasons	3.52	1.46	14.45	5.91
Support	3.63	1.43	15.16	5.98
Organization	3.55	1.52	14.85	6.04
Sentence variation	3.48	1.47	14.23	5.89
Diction	3.57	1.55	15.11	6.12
Grammar errors	3.58	1.52	15.05	6.07
Example	3.61	1.52	15.32	6.11

Table 1 shows a description of the mean, standard deviation, mean square, and standardized Z. Average opinion/position of 3.75 is the highest average value compared to other variables (SD = 1.56). The mean square error value is the average difference between the measurement and forecasting values. The lowest average measurement error is the sentence variation variable (14.23). While the highest average measurement error is the opinion/position variable (16.41). Table 2 compares the response values or statements that are not expected from each variable or criterion.

Table 2. Total instances of unexpected responses by criteria

Items	Instances	% total instances
Support	27	0.023
Organization	37	0.031
Sentence variation	41	0.034
Diction	39	0.033
Grammar errors	35	0.029
Example	31	0.026
Total unexpected responses	210	0.175
Total response	1,200	

Overall, there were 210 unexpected responses from all variables, and the most widely found in the sentence variation variable was 41 responses. Meanwhile, the lowest number is found in the support variable, with 27 responses. In total, there are 210 unexpected responses, or around 0.17%. Generally, it is very small (< 1%). Table 3 shows the writing test variable, the percentage of students who score more than 3 is 63%, while students who score less than 3 are only 16%. This shows that the writing test ability of the students tested in this study is quite good.

Table 3. Average candidate ability measures

Category score	Times category used	%	Cumulative %
1	6	6	6
2	10	10	17
3	20	21	38
4	15	16	53
5	18	19	72
6	27	28	100

3.2. Chi-square test

Chi-square test is a non-parametric comparative test performed on two variables where the data scale of the two variables is nominal. If of the two variables, there is one variable with a nominal scale, a Chi-square test is carried out with reference to that the test at the lowest degree must be used. The Chi-square test is the most widely used non-parametric test. The results of the Chi-Square test are presented in Table 4.

Table 4. Chi-square test results

Items	Group	N	Observed proportion	Expected proportion	X^2	Φ	Sig (2-tailed)
Argument	Agree	80	0.80	0.50	87.756	92.889	0.000
	Disagree	20	0.20				
	Total	100	1.00				
Topic	Agree	80	0.80	0.50	89.304	92.889	0.000
	Disagree	20	0.20				
	Total	100	1.00				

In the argument item, 20% of the research subjects stated that they did not agree with the argument variable statements related to the writing test. In other words, 1 out of 5 respondents disagreed with the statement in the argument variable. When viewed from the $X^2{}_{yates\ value}$ of 87.756 with a significance of 0.000, it can be concluded that the item argument has a relationship with the writing test variable.

On the topic item, 20% of the research objects stated that they disagreed with the topic variable statements related to the writing test. In other words, 1 out of 5 respondents disagreed with the statement in the topic variable. Therefore, based on the $X^2{}_{yates\ value}$ of 89.304 with a significance of 0.000, it can be concluded that the topic item and the writing test variable are related.

The data analysis results show that the coefficient of determination (R) between the writing test and reading and writing is quite high at 0.997 (99.7%). This indicates that reading and writing are able to explain the writing sample variability by 99.7%. Therefore, this strong correlation indicates that the ability to read and write has a significant influence on the variability present in the writing samples.

The results in the final result adjusted 1 and final result adjusted 2 categories are almost similar to those in the final result category; each coefficient of determination value is 0.995 (99.5%) and 0.991 (99.1%). Furthermore, the coefficient of determination (r^2) between the writing test and listening and speaking is also high at 0.995 (99.5%), indicating that the listening and speaking variability explains the writing sample variability by 99.5%. The results in the final result adjusted 1 and final result adjusted 2 categories are similar to those in the final result category, with the coefficient of determination of 0.991 (99.1%) and 0.985 (98.5%), respectively. Comparing the two, the coefficient of determination of the reading and writing variables is slightly better than the listening and speaking.

Table 5 presents the coefficient of determination between writing samples categorized by course. This analysis allows for a deeper understanding of how different courses may influence the writing proficiency of students. Table 6 displays the correlation between the writing test and courses. This analysis provides an overview of the extent to which the writing test correlates with specific subjects within the curriculum.

Table 5. Coefficient of determination between writing sample by course

Course	Result	Items	Level 1
Reading and writing	Final result	r^2	0.997
		n	100
	Final result adjusted 1	r^2	0.995
		n	81
	Final result adjusted 2	r^2	0.991
		n	60
Listening and speaking	Final result	r^2	0.995
		n	100
	Final result adjusted 1	r^2	0.991
		n	81
	Final result adjusted 2	r^2	0.985
		n	60

Table 6. Correlation of writing test and course

Course	Items	Reading comprehension	Sentences skills	Total accuplacer
Writing Sample	r	0.995	0.995	0.997
	p	0.000	0.000	0.000
	n	100	100	100

The correlation coefficient between the writing sample and reading comprehension, sentences skills, and accuplacer total are all above 95% (strong correlation). This result means that the writing test is closely related to reading comprehension, sentence skills, and total accuplacer. Therefore, of the three variables, when compared, the total accuplacer variable has the greatest correlation with the writing test. In other words, the total accuplacer variable has the highest correlation with the writing test.

3.3. Wilcoxon test

The wilcoxon signed rank test is a nonparametric statistical test used to assess the significance of the contrast between two sets of paired data on an ordinal or interval scale, particularly when the data is not normally distributed. Especially when dealing with smaller sample sizes or data that does not meet the assumptions of parametric tests. Wilcoxon test results on the writing test variable are displayed in Tables 7 to 9 offer insight into the specific statistical comparisons between the paired data of the writing test variable, shedding light on any significant differences identified through this nonparametric analysis.

The descriptive statistics in Table 7 show the mean, standard deviation, minimum, and maximum values of each data group (pretest and posttest). The posttest mean is 3.5606, smaller than the pretest (3.6630). The difference in means between the post-test and pre-test indicates a decrease in the average scores of the writing test from the initial assessment to the subsequent test. This decline suggests a potential change or reduction in performance between the two tests. Table 8 presents the ranks test results. This statistical analysis is conducted to compare the relative performance of participants across different conditions or groups.

Table 7. Descriptive statistics

	Descriptive statistics				
	N	Mean	Std. Deviation	Minimum	Maximum
Pretest	100	3.6630	1.55115	1.20	6.00
Posttest	100	3.5606	1.49596	1.03	5.89

Table 8. Ranks test

Ranks				
		N	Mean rank	Sum of ranks
Posttest - pretest	Negative ranks	100[a]	50.50	5050.00
	Positive ranks	0[b]	0.00	0.00
	Ties	0[c]		
	Total	100		

Note: a. Posttest < pretest, b. Posttest > pretest, c. Posttest = pretest

The wilcoxon signed rank test results in Table 9 show the Z value of -8.718 (p = 0.000), meaning the pretest and posttest groups on the variable writing test are significantly different. The low Z value and the very low p-value indicate a highly significant difference between the pretest and posttest groups concerning the writing test variable. These significant results affirm a noticeable disparity in performance or scores between the initial assessment and the subsequent one, suggesting a substantial change or improvement. Therefore, the evidence from the wilcoxon signed rank test in Table 9 unequivocally supports the assumption that there is a significant difference in writing test scores before and after the intervention or testing period.

Table 9. Test statistics

Test statistics	
	Posttest–Pretest
Z	-8,718 [b]
Asymp. Sig. (2-tailed)	0.000

Note: a. Wilcoxon signed ranks test, b. Based on positive ranks.

3.4. Discussion

This study aimed to examine the relationship between writing tests and assessment writing through argumentative writing. These research findings may help teachers' knowledge on the influence and relationship between writing tests and assessment writing have on students' writing scores. Understanding how writing tests and evaluative criteria, particularly in argumentative writing, contribute to assessing students' writing abilities is crucial. The potential findings of this research can enhance educators' comprehension of how specific elements within writing assessments influence students' overall writing scores. By delineating a clearer relationship between these factors, educators can adapt teaching strategies and assessment approaches to better align with students' needs, ultimately improving their writing abilities.

RQ1: relationship between writing tests and assessment writing through argumentative writing

The Chi-square test results in Table 4 indicate the relationship between certain variables and the writing test. It was found that 20% of the research subjects disagreed with the statements in the argument and topic items related to the writing test. Data analysis also revealed a high coefficient of determination between the writing test and reading, as well as writing, and between listening and speaking, indicating a significant influence of reading and writing abilities on the variability in writing samples. The final results in the adjusted categories showed nearly identical coefficient of determination values, with the reading and writing variables slightly outperforming listening and speaking.

The test results in Table 5 show the coefficient of determination between the writing sample or argumentative writing on the course, the correlation between writing and reading, and writing and listening and speaking, with reading having a stronger relationship with the writing test than listening. The coefficient of determination (r^2) between the writing-reading and writing tests is high (0.997 or 99.7%). This demonstrates that reading and writing may account for 99.7% of the variability in the writing sample. However, the coefficient of determination (r^2) between the writing and listening-speaking tests is also high (0.995 or 99.5%). It is slightly better than the listening and speaking variables. The results of this study are supported by Plakans [9], stating that the writing-only task and the reading-to-write task elicited different compositional behaviors, with the latter lead to a more interactive process. It means that the writing-only task and the reading-to-write task have a strong relationship with the writing test and are carried out in an interactive process with various compositional behaviors. The research results reveal that the correlation between writing tests and assessments writing through argumentative writing is significant, as shown in Table 6. The correlation between the writing test and the course, namely the reading comprehension and sentence skills items, is above 95%, indicating a strong relationship. In other words, reading comprehension and sentence skills items correlate highest with the writing test.

The results of the questionnaire administered to the students in this study show that the argument and topic items have a substantial relationship to the writing test. This finding is supported by the questionnaire results, showing that 20% of the study objects disagreed with the statement of the argument variable associated with the writing test on the argument item. The argumentation item does, however, associate with the writing test variable, as evidenced by the X^2 score of 87.756 (p = 0.000). Concerning the topic item, 20% of the research objects said they disagreed with the topic variable's assertion about the writing test. As a result, the topic item is linked with the writing test variable, as evidenced by the X^2 value of 89.304 (p = 0.000). This finding agrees with Beck and Jeffery [15], who said that argumentative writing is an important pedagogical genre in writing that should be included in high-risk assessments.

4. CONCLUSION

Research findings reveal a close relationship between writing tests and assessment writing. Overall, both variables have the same strength and are equally important and related. The questionnaire results given to participants reveal that the items of argument and topics have a significant relationship with the writing test. Furthermore, in the correlation test between reading and writing and listening and speaking, the outcome of the writing assessment reading has a stronger relationship with the writing test compared to listening. This study has several limitations, some of which could be improved upon in future research. Future research should use a wider component and a larger number of participants. The consequences of this research for language learning and teaching are numerous.

ACKNOWLEDGEMENTS

The Universitas Pahlawan Tuanku Tambusai, Indonesia, funded the project. The researchers express their gratitude to the University for providing funding for this project, as well as to the anonymous reviewers and editors for their valuable suggestions and comments during the preparation of this article.

REFERENCES

[1] R. Callahan, L. Wilkinson, and C. Muller, “Academic achievement and course taking among language minority youth in u.s. schools: effects of esl placement,” *Educational Evaluation and Policy Analysis*, vol. 32, no. 1, pp. 84–117, 2010, doi: 10.3102/0162373709359805.

[2] M. Hodara, “The effects of english as a second language courses on language minority community college students,” *Educational Evaluation and Policy Analysis*, vol. 37, no. 2, pp. 243–270, 2015, doi: 10.3102/0162373714540321.

[3] N. Behizadeh, “Mitigating the dangers of a single story: creating large-scale writing assessments aligned with sociocultural theory,” *Educational Researcher*, vol. 43, no. 3, pp. 125–136, 2014, doi: 10.3102/0013189X14529604.

[4] W. Condon, “Large-scale assessment, locally-developed measures, and automated scoring of essays: fishing for red herrings?,” *Assessing Writing*, vol. 18, no. 1, pp. 100–108, 2013, doi: 10.1016/j.asw.2012.11.001.

[5] P. A. Moss, “Validity in high stakes writing assessment: problems and possibilities,” *Assessing Writing*, vol. 1, no. 1, pp. 109–128, 1994, doi: 10.1016/1075-2935(94)90007-8.

[6] L. F. Bachman, “Building and supporting a case for test use,” *Language Assessment Quarterly*, vol. 2, no. 1, pp. 1–34, 2005, doi: 10.1207/s15434311laq0201_1.

[7] B. Bridgeman, D. Powers, E. Stone, and P. Mollaun, “TOEFL ibt speaking test scores as indicators of oral communicative language proficiency,” *Language Testing*, vol. 29, no. 1, pp. 91–108, 2012, doi: 10.1177/0265532211411078.

[8] G. Fulcher, “Assessment in english for academic purposes: putting content validity in its place,” *Applied Linguistics*, vol. 20, no. 2, pp. 221–236, 1999, doi: 10.1093/applin/20.2.221.

[9] L. Plakans, “Comparing composing processes in writing-only and reading-to-write test tasks,” *Assessing Writing*, vol. 13, no. 2, pp. 111–129, 2008, doi: 10.1016/j.asw.2008.07.001.

[10] A. Cumming, L. Grant, P. Mulcahy-Ernt, and D. E. Powers, “A teacher-verification study of speaking and writing prototype tasks for a new toefl,” *Language Testing*, vol. 21, no. 2, pp. 107–145, 2004, doi: 10.1191/0265532204lt278oa.

[11] L. Brooks and M. Swain, “Contextualizing performances: comparing performances during toefl ibttm and real-life academic speaking activities,” *Language Assessment Quarterly*, vol. 11, no. 4, pp. 353–373, 2014, doi: 10.1080/15434303.2014.947532.

[12] S. C. Weigle and E. Friginal, “Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: effects of language background, topic, and l2 proficiency,” *Journal of English for Academic Purposes*, vol. 18, pp. 25–39, 2015, doi: 10.1016/j.jeap.2015.03.006.

[13] M. Rahimi and L. J. Zhang, “Effects of task complexity and planning conditions on l2 argumentative writing production,” *Discourse Processes*, pp. 1–17, 2017, doi: 10.1080/0163853X.2017.1336042.

[14] D. Lazere, “Into the field: sites of composition studies and the powers of literacy: a genre approach to teaching writing,” *JAC*, vol. 15, no. 1, pp. 176–182, 2016, [Online]. Available: https://www.jstor.org/stable/20866018

[15] S. W. Beck and J. V. Jeffery, “Genres of high-stakes writing assessments and the construct of writing competence,” *Assessing Writing*, vol. 12, no. 1, pp. 60–79, 2007, doi: 10.1016/j.asw.2007.05.001.

[16] M. A. Tabari, “The effects of planning time on complexity, accuracy, fluency, and lexical variety in l2 descriptive writing,” *Asian-Pacific Journal of Second and Foreign Language Education*, vol. 1, no. 10, pp. 1–15, 2016, doi: 10.1186/s40862-016-0015-6.

[17] F. A. Al-Haq and A. S. E. A. Ahmed, “Discourse problems in argumentative writing,” *World Englishes*, vol. 13, no. 3, pp. 307–323, 1994, doi: 10.1111/j.1467-971X.1994.tb00318.x.

[18] S. Link, M. Mehrzad, and M. Rahimi, “Impact of automated writing evaluation on teacher feedback , student revision , and writing improvement,” *Computer Assisted Language Learning*, pp. 1–30, 2020, doi: 10.1080/09588221.2020.1743323.

[19] L. Plakans, A. Gebril, and Z. Bilki, “Shaping a score : complexity , accuracy , and fluency in integrated writing performances,” *Language Testing*, pp. 1–19, 2016, doi: 10.1177/0265532216669537.

[20] T. Park, “An investigation of an esl placement test of writing using many-facet rasch measurement,” *Teachers College, Columbia*

University Working Papers in TESOL & Applied Linguistics, vol. 4, no. 1, pp. 1–21, 2005.

[21] K. Wijekumar *et al*., "The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material," *Red Writ*, pp. 1–27, 2018, doi: 10.1007/s11145-018-9836-7.

[22] J. Ong and L. J. Zhang, "Effects of task complexity on the fluency and lexical complexity in efl students ' argumentative writing," *Journal of Second Language Writing*, vol. 19, no. 4, pp. 218–233, 2010, doi: 10.1016/j.jslw.2010.10.003.

[23] J. Li, S. Link, and V. Hegelheimer, "Rethinking the role of automated writing evaluation (awe) feedback in esl writing instruction," *Journal of Second Language Writing*, vol. 27, pp. 1–18, 2015, doi: 10.1016/j.jslw.2014.10.004.

[24] I. Lee, P. Mak, and R. E. Yuan, "Assessment as learning in primary writing classrooms : an exploratory study," *Studies in Educational Evaluation*, vol. 62, no. April, pp. 72–81, 2019, doi: 10.1016/j.stueduc.2019.04.012.

[25] S. Raoofi and Y. Maroofi, "Relationships among motivation (self-efficacy and task value), strategy use and performance in l2 writing," *Southern African Linguistics and Applied Language Studies*, vol. 35, no. 3, pp. 299–310, 2017, doi: 10.2989/16073614.2017.1391706.

[26] L. Wang, I. Lee, and M. Park, "Chinese university EFL teachers ' beliefs and practices of classroom writing assessment," *Studies in Educational Evaluation*, vol. 66, pp. 1–11, 2020, doi: 10.1016/j.stueduc.2020.100890.

BIOGRAPHIES OF AUTHORS

Masrul Masrul is the lecturer of Universitas Pahlawan Tuanku Tambusai. He is Master in English Education and he has completed his Doctoral degree in Universitas Negeri Padang. He taught at Department of English Education and he is expert at writing, reading, assessment, and evaluation of English education. He has written several four articles related to the assessment of students' reading and writing and five authors have cited his article. Moreover, he is also active on writing program which focus on research and writing scientific article. He can be contacted at email: masrulm25@gmail.com.

Santi Erliana is active lecturer at Institut Agama Islam Negeri Palangkaraya. She has good English ability both productive (writing, speaking) and receptive skill (listening, reading). She is expert at reading, CLIL, English as medium of Instruction (EMI) and writing. He is also concerned on improving literacy and reading comprehension. She has been active writer since 2016, she has published 22 researches and her research has been cited by 23 authors. She can be contacted at email: santi.erliana@iain-palangkaraya.ac.id.