Advances in Social Science, Education and Humanities Research, volume 174
First Indonesian Communication Forum of Teacher Training and Education Faculty Leaders International Conference on Education 2017 (ICE 2017)

Developing Argumentative Essay Writing Test

Anshari Syafar
Universitas Tadulako
Palu, Indonesia
syafaranshari@gmail.com

***Abstract*—This article is to develop argumentative writing test needs thoroughly analysis to have appropriate test which possess high quality of validity and reliability. Writing test which has been developed in this study is argumentative writing test which is so-called Test Development Plan. The test was designed and developed on the basis of general objective of the course "Writing IV" of English Study Program Universitas Tadulako. Aspects of the test were developed systematically including critical revision, reformulation, and justification of the test format to come up with general and specific objectives of the test draft. Then, the draft was subsequently developed by setting up criteria, prompt, and scoring guide for designing and producing an appropriate format of argumentative essay writing test. The developed fomat test was tried out by asking students to write argumentative essays based on the topics assigned. The students' essays were graded by two raters to have scores for analyzing its validity and reliability. After analyzing scores given by the two raters, the developed test indicates to have high validity and reliability. Furthermore, the test has fulfilled standard of argumentative essay test and it is applicable for measuring and grading students' skill in writing argumentative essay.**

Keywords—argumentative essay test; test format; scoring guide; test validity; realibility, prompt, grading

I. INTRODUCTION

Theoretical basis of language testing and evaluation are due to the inception of many language approaches and methods applied in language teaching and encounters multifaceted problems in evaluating and grading capability of students in language skill and competence**.** Moreover, as interaction among people from different countries increase, language education becomes an essential element in improving the effectiveness of communication. It is commonly recognized that an age of multilingualism has arrived. Thus, over several decades, language educators have developed many different teaching approaches, methodologies, and techniques to enhance the effectiveness of foreign/second language teaching and learning. However, language teachers encounter many problems to evaluate and grade language competence and skills of students both productive and receptive skills.

In line with, the role of testing or evaluation, in general context is used in learning, certification, determination of accountability, ranking, monitoring and prediction [1]. However, we know that language testing can play many important roles in helping people to evaluate; it is reasonable to assume that language testing can be used as an effective tool to enhance teaching directly or indirectly. Conversely, the real potential of language testing as an effective tool to improve learning is seldom realized. Therefore, evaluation in language teaching can be defined as an endeavor to collect information for making decision on educational issues and policies [2].

In past decades, language educators and testers spent a lot of time and energy in attempts to find valid tests to meet the needs of the language education field. A common focus was on the construct, language competence, since gauging competence is a primary purpose of language testing—measuring language ability in a reliable and valid ways so that success in meeting educational objectives can be understood. This can be witnessed throughout the history of language testing. Therefore, [3] states that writing ability may, broadly, be thought of as having four contributing factors. First, writers require familiarity with the content that they are to write about and the ability to reason with this content so as to present a coherent and convincing account to their audience. Second, writers require familiarity with the genre of the text that they are to produce, and understanding the particular registers and structures that are required by the community for whom they are writing. Third, writers require meta-cognitive skills for managing the interaction between content and expression and, thus, for developing argument. Finally, writers require an overarching method-of-working to provide a framework within which the detailed goal setting and decision making associated with writing.

Correspondingly, [4] states that testing language skill is difficult, but testing writing of students of English as a Second Language poses, two major problems. The first is making decisions about the matter of control, objectively of the evaluation, and naturalness in the writing test. The second major problem is that, if the test is done in a way that cannot be graded objectively, it is necessary to develop a scale that make grading as objective as possible. The ability to write involves grammatical and lexical abilities, mechanical ability, stylistic and organizational skills, and the ability to judge whether something is appropriate.

Meanwhile, almost authors agrees that common tasks for writing tests included: (1) gap filling; (2) form of completion; (3) making correction; (4) letter writing; (5) dictation and dicto-comp; (6) grammatical transformation; (7) short answer and sentence completion; and (8) essay writing [2-9]. Any chosen task should be evaluated for its relevance to the

student's eventual use of the language. When testing students at the intermediate and advanced levels, test developer or maker must consider the instructions, the choice of topics, the choice of tasks, and the level of difficulty and time allowed. All these considerations must go into making a test that is appropriate for learner, and then the test developer must attempt to ensure that marking the test, which will always be at least somewhat subjective, is as objective as possible [2, 4, 5, 10].

In the context of writing argumentative essay, [11] narrates that studies demonstrating that argumentative essays, for instance, are syntactically more complex than narrative or descriptive essays are not very interesting in the absence of some way of objectively classifying those modes. In most published studies, researchers avoid the problem of classification by assuming that the "mode" of an essay is determined by the intentions of the writer of the stimulus topic. That is, a set of essays written in response to a topic calling for, say, argumentation is assumed to be uniformly in an argumentative mode, so that the characteristics of the essays can be reported as representative of arguments.

II. METHODS

Developing argumentative writing test was part of project of test development plan (TDP) to produce test which had traits of good validity and reliability to apply in grading students' works on writing skill. Futhermore, the first plan to carry out was to set up general description of the TDP outlining the background, rationale, process and product of the project. In this context, establishing the descriptors on such an analysis would contribute to the empirical value of the scale [12]. Then, the stages to take in developing argumentative essays test consisted of planning, developing, testing validation, and trying out the test. In the stage of planning the test, some aspects had to take into account including general and specific objective of the course—Writing IV, general and specific objective of the test, kind and format of the test, table of specification, and source of test material of the test. Developing stage of the test covered process of developing test item, proving answer sheet and deciding scoring and grading criteria. In the test validation stage, test developer reviewed the entire test item including criteria, prompt and scoring guide that had been set up in the test development stage. The last stage was to tryout the developed test which employed implementation, grading and scoring, and analysis of students' scores graded for their argumentative essays.

In line with, general objective of the course—Writing IV states that by the end of the semester, the students are expected to be able to write thoroughly various English essay e.g., description, narration, expository, argumentative and writing scientific paper or report [13]. With the general objective of the course, the writer takes it as the baseline to develop general and specific objective of the test. General objective of the test states that this test is intended to measure students' ability to write argumentative essay. Meanwhile, specific objective of the test narrates that the test is aimed to measure students' ability in writing an argumentative essay which covers the following competences: (1) to write logical development of argumentative ideas in which one is taking position to agree or disagree on a certain topic by giving reasons based on the order of importance; (2) to write argumentative essay with good organization in terms of clarity, unity and relevance and supporting details e.g., reasons, facts and examples and (3) to write argumentative essay with the use of appropriate language and mechanics.

In line with the test plan, table of specification is also developed to portray proportion of criteria assigned on each competence expected to be performed by students, the test-takers. Three criteria are assigned as ability elements to be performed by students in writing argumentative essay. These criteria are content which is weighted 40%, that is, to write logical development content of argumentative ideas in which one is taking position to agree or disagree on a certain topic by giving relevance reasons and supporting details e.g., facts and examples based on the order of importance. The organization is weighted 30%, that is, to write argumentative essay with good organization in terms of clarity, unity, and coherence. Language use is weighted 30%, that is, to write argumentative essay with appropriate language use—syntactical rules and word choices.

A. Test Validation

Ref. [14] states that an argument ensues when people disagree about something in which one side gives an opinion and offer in support of it, and the other side gives a different opinion and offers reasons in support of his or her stand. Thus, an argumentative essay is to take a stand or opinion that attempts to change reader's mind, so that the topics of the assay assigned should be interesting to capture the writer's curiosity, adaptability to take a stand or opinion, and novelty to cope with the current phenomena related to cultural, social, or educational issues. Underlying the abovementioned statements, the topics assigned in this TDP are chosen because of having qualities of interesting, adaptability, and novelty. These topics are also discussed in two reference books of writing IV [14, 15] used by the lecturer to teach the course. Therefore, the choice to assign some topics by test developer intends to reduce multiple-interpretation and subjectivity of criteria assigned in scoring, since writing test is subjective test. [2] states that essay test generally refer to subjective test in which it can only be scored subjectively.

In the point of view, test validation covers three stages: (1) review of the entire test; (2) validity: (3) reliability. In reviewing entire test, test developer reviewed some refernces as the stage of developing prompt, scoring guide and identification of answers. The second factor is validity of the test developed. The test developer interpreted validity as developing criteria of the topic, that is, argumentative essay to state skills or abilities expected to be performed by the test-takers. The third factor was reliability of the test being developed. In order to have reliability of the argumentative essay writing test, the test developer adapts the procedure

applied by [2] using formula of Pearson Product-Moment Coefficient Correlation to correlate the raters' scores—rater1 and rater2 as follows:

$$r - xy = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N\, S_x\, S_y} \quad (1)$$

Where:
r - xy : Coefficient Correlation Pearson Product-Moment
$\sum$: Sum
X : Individual score of test X
Y : Individual score of test Y
$\bar{X}$: Mean score on test X
$\bar{Y}$: Mean score on test Y
S_x : Median score on test X
S_y : Median score on test Y
N : Total number of test-takers

To have more accurate reliability estimation, [2] suggests that: (1) raters should have relatively the same competence on the topics tested; (2) they score students' essay based on the same scoring guide and criteria assigned; and (3) final score should be determined by calculating mean score from each rater without having extreme score differences, that is, beyond the maximum score that has been assigned. The maximum scores ranges from minimum 10 to maximum 40. In this case, both raters fulfill the criteria to be scorer because they are both teacher of writing courses; they are master graduate in related subject matters, they will use the same scoring guide to score students' essays. Hence, there is no doubt about the competence of raters in scoring students' essays based on the criteria assigned as reflected in table of specification and scoring guide.

B. Scoring and Grading

In essay writing, there are several approaches that can be used to score writing essay: such as holistic scoring; analytic scoring; primary trait scoring; and multi-trait scoring [2, 5, 6]. Holistic scoring is one that is based on a single, integrated score of writing behavior. Since holistic scoring requires a response to the writing as a whole, respondents are unlikely to be penalized for poor performance on one lesser aspect (e.g., grammatical ability [6]. He further explains that analytic scoring calls for the use of separate scales, each assessing a different aspect of writing—for example, content, organization, vocabulary, grammar and mechanitcins. Correspondingly, [12] elucidates that establishing the descriptors on such an analysis would contribute to the empirical value of the scale. Rating scale was set up in the TDP based on the scoring and grading system of the course.

The rating scale allotted in the scoring guide is *"4 to 1"*. This rating scale is firstly applied to score on each level of students' performance in their essays as described in the scoring guide. Then, scores gained from rating scale 4 to 1 is then weighted to have equivalence score according to the criteria assigned, that is, 40% on content and 30% on organization and language use. Score 4, for instance, is weighted 40% to get equal score 16 on content and score 4 is weighted 30% to have equal score 12 on organization and language use, etc. With this weighted score, it will generate maximum score 40 and minimum score 10 that is obtained from three criteria assigned in scoring guide. Then, individual scores and group scores—mean, mode, median, minimum and maximum scores, and standard deviation is used to describe the tendency of students' score distribution, score variability and to test validity and reliability of the test that has been developed and tried out.

III. Results and Discussion

The developed test was implemented in order to see the correspondence of the test with students argumentative essays. Therefore, 25 students of English Study Program were asked to write argumentative essay based on the prompt. Each student was expected to produce argumentative essay in about 450-500 words. The students' argumentative essays were rated by two raters to have scores for analysis of validity and reliability of the test that had been developed by test developer. Therefore, their essays had to be sorted and filed to hide students' identity on their work before the two raters grade their essays. Test developer first filed their identity in computer erased them by using correction-pen. Then, he changed students' identity into code "Respondent 1 to 25 or R1 to R25" intending to avoid bias on familiarity with the name by raters. Test developer firstly scored students' essays, he took those essays back to the teacher attached with the scoring sheet given respectively code R1 up to R25, so that the teacher did not need to have her own paper to put on her scoring.

A. The Scores of Students' Essays

The following is general description of scores were elicited from two raters. The two raters were rater1 (R1), test developer and rater2 (R2), the teacher of Writing IV. In table 1, the list of scores of students' argumentative essays reflects the obtained scores from rater1 and rater2, the difference of score between the two raters, and mean scores of both individual and groups. Since the tolerable maximum difference is set up at 6, there are two scores on rows R18 and R20 in table 1 gained from raters going above the tolerable differences. Thus, these two scores should be corrected to come up with tolerable deviation; then we needed third rater for adjusting the scores. Total scores obtained from the third rater for R18 is = 22 and for R20 = 33. These scores have been applied to add up the three scores obtained from (rater1, rater2 and the third rater) and divided by three to get mean score. These average score of R18 and R20 can be seen in table 1 below.

TABLE I. THE SCORES OF STUDENTS' ARGUMENTATIVE ESSAY TEST

Students	Score of Rater 1	Score of Rater 2	Diffrences	Mean
R1	33	30	3	31.50
R2	23	27	3	25.00
R3	33	37	4	35.00
R4	23	23	0	23.00
R5	20	20	0	20.00
R6	17	17	0	17.00
R7	33	33	0	33.00
R8	23	23	0	23.00
R9	30	27	3	28.50
R10	23	23	0	23.00
R11	26	20	6	23.00
R12	23	27	4	25.00
R13	24	27	3	25.50
R14	27	30	3	28.50
R15	20	17	3	18.50
R16	33	33	0	33.00
R17	30	30	0	30.00
R18	16	23	7	20.30
R19	16	20	4	18.00
R20	30	37	7	33.30
R21	26	30	4	28.00
R22	27	30	3	28.50
R23	33	36	3	34.50
R24	36	37	1	36.50
R25	26	27	1	26.50
Total score				668.10
Mean score				26.72
Maximum score				36.5
Minimum score				17
Median				27.25
Mode				23
Standard deviation				5.69

B. *The Tendency of Students' Score Distribution*

To make easier of calculating and estimating students score, the test developer applied computer program Microsoft Excel and SPSS 15.01 for Window to help him summarize the row scores as reflected in Table 1. The table reveals the tendency of students' scores on argumentative essay writing test. The average is obtained by summing up (∑) the total score of argumentative essay gained from two raters (X), which is then divided by number of respondent (N). The formula is:

$$\bar{X} = \frac{\sum X}{N}$$

$$\bar{X} = \frac{668.1}{25} = 26.72$$

The mean score (26.72) obtained from the test indicates that the scores are distributed along the scale. This argument is based on the rank of individual score in which the mean is exactly at the balance point of the score distribution. This fact can be noticed if we sort the scores in rank order from highest score to the lowest score. We could see that 12 scores fall above the mean and 12 scores fall below the mean. These scores indicate the level of students' performance in writing argumentative essay which can be classified fairly higher. Reference states that the higher the mean score is, and so, the higher the level of individual score obtained by test-takers in the group. On the contrary, the lower the mean score is, and so, the lower the level of individual score obtained by test-taker in the group.

In line with central tendency of the score distribution, the median of students' score distribution tendency indicates in all the scores as being arranged in ranking order. Since median is the score which is at the centre of the distribution, then half of the scores are above the median and half below. However, if the number of the score is odd, the median is in the middle score; if the number of score is even, used the midpoint between the two middle scores as the median [2]. In this case, the students' scores are even and the midpoint between the two middle scores is: (26.50 + 28 ÷ 2 = 27.25). Hence, the median score of students' argumentative essay is 27.25, which takes place in between scores 26.50 and 28.

Correspondingly the mode tendency of students' score as distributed based on the two raters, total scores and average score. Rater1 has tendency mode score 23, while rater2 has tendency to have two mode scores—27 and 30. The total mode scores obtained from the two raters is 46, and the average mode score is 23. This mode score indicates that score gained from rater1 tends to be unimodal and rater2 has bimodal mode score. However, score 23 is the mode of group performance in argumentative essay writing test and this mode tends to be unimodal mode.

C. *Score Variability*

It is different from those of central tendency analysis, which is intended to see the tendency of individual score as a group of test-takers in general. Score variability analysis is intended to describe how obtained individual score of test-taker varies and differentiates entirely [2]. To know the score variability, we need to calculate it by using three kinds of indicators such as: (1) score range; (2) standard deviation and (3) variance.

Furthermore, the formula used to calculate scores of students' argumentative essay is score range (R), that is, the different between the highest score (H) and the lowest score (L) toward score obtained from the two raters of students' essays. To have more details information of score of students, test developer presents in the form of table involving the two raters' score, total score and average score on top row and column. While on the left columns we can see level on the top followed by highest score, lowest score and score range respectively. The range score is portrayed based on the scores obtained from rater1, rater2, total score—the summed score from rater1 and rater2, and average score—the score gained by dividing by two of the total score. The formula is suggested as follows: R = H – L

In order to portray score range of students' argumentative essays in detail information, test developer sets it up in the following table presentation. Table 2 shows score range from rater1 or R1 (test developer), rater2 or R2 (teacher), range of

total score, and range of average score. From the four kinds of score ranges is presented in Table II, it indicates that the range of score gained from R1 and R2 indicating there is no significantly different among the three kind of score ranges.

TABLE II. SCORE RANGE OF STUDENTS' ARGUMENTATIVE ESSAY

Levels	Score of Rater 1	Score of Rater 2	Tota Scores	Mean
Highest score	H36-	H37-	73-	36·5-
Lowest score	L16	L17	34	17
Score Range	R=20	R=20	=39	=19.5

Standard deviation (S) portrays level of score variation that is strongly and mostly used in descriptive statistic calculation [2]. The standard deviation of students' argumentative essay scores can be calculated by applying the following formula:

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$S = \sqrt{\frac{778.72}{N}}$$

$$= \sqrt{31.14}$$

$$= 5.58$$

The standard deviation of students' scores on their argumentative essay is 5.58. This figure indicates variability of students' score distribution around the mean. The standard deviation obtained, according to test developer, is greater. Therefore, it can be interpreted that the students' scores are spreading variably from the central point in the distribution.

In line with, the procedure of calculating standard deviation can be done in a different method, that is, to calculate the weight of variance which is defined as square of standard deviation (S^2), or square of mean deviation [2]. In this case, the variance of student argumentative essays' score can be obtained by the use of variance calculation formula as follows:

In most statistical analyses, the variance is used to measure of variability. Variance is the sum of the squared deviation scores divided N–1 [16]. According to [2], variance is defined as variability of factors that can affect performance of test-takers in doing the test. Because of the interference of these factors, too deviations can affect test-takers' work performance due to the influence of variances, such as: place of testing, procedure of testing, procedure of scoring, developing and designing test format and item, and state condition of test-takers when taking the test. All the said variances could happen when this test was being tried-out.

D. Analysis of Test Validity

To analyze the validity of test that had been tried out, test developer underlies to construct validity as assigned in TDP. Construct validity in TDP suggest that students have to write an argumentative essay by taking one opinion or position from one of the four topics provided. Their assay should convey argumentative idea with logical development starting from thesis statement as the first paragraph, order of importance of reasons, and conclusion paragraphs. The essay has to be also developed with good organization—unity, clarity, and coherence; supporting details—reasons, facts, and examples; appropriate use of language grammar—correct structure (present tense or past tense), words choice; and mechanics—punctuations, capitalizations, and spelling.

In order to have validity of the test, it is needed to have criteria for scoring students' essay. The criteria are used to assign competences that must be demonstrated by the test-takers in the test. If test takers can demonstrate the competences as good as assigned in the criteria, then the test is said to have validity. Reference [2] states that the aspect of validity that is measured through description and reasoning can only be verified through qualitative approach by labeling the level of validity as "high, moderate, and low validity." Therefore, analysis of test validity is done through presenting score of students' argumentative essay in empirical data and thoughtful descriptions. Three related factors are used to validate argumentative essay writing test including the prompt, topic and criteria assigned. The prompt with clear instructions elicits data as designated in the instruction. In the prompt for instance, students are assigned to write argumentative essay consisting of 4 to 5 paragraphs with 350-400 words. This assigned prompt corresponds with the content of students' argumentative essays. The average number of paragraph the students produced in their essay is 4.56 or (four point fifty six) paragraphs in each essay, while the average number of words is 347.32 or (three hundred forty seven point thirty two words) in each essay.

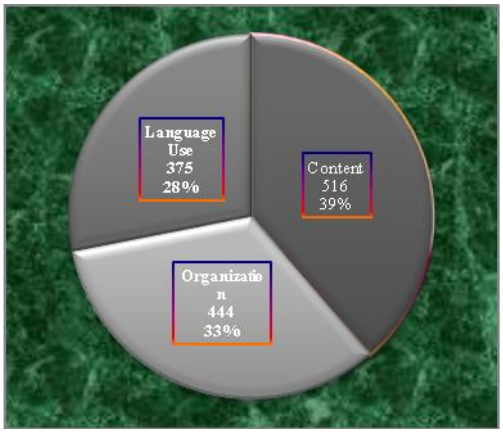

Fig. 1 Percentage of students' scores on their argumentative essay

In relation to the criteria assigned in the table of specification and scoring guide—content, organization and language use, scores of students' argumentative essay can also be used to analyze test validity based on the score percentage given by the two raters. Analysis is started from criteria assigned and weighted in percentage (content 40%,

organization 30%, and language use 30%). Then, the percentage of criteria is matched with the percentage of score obtained from the two raters. If the scores given by rater1 and rater2 is added up as reflected as in Fig. 1, then we could have the value of percentage closely the same amount of the percentage as assigned in the criteria. Content for example, is weighted 40% and scores of students' argumentative essay obtained is 39%. It is only 1% splitting from assigned criteria. While organization which is weighted 30% goes above 3% from assigned criteria with 33% score obtained from students' argumentative essay. Language use is also weighted 30% lost 2% from assigned criteria with 28% score of students' argumentative essay obtained.

Test developer uses this figure to analyze the validity of the test, because he bases his understanding on the concept of a testing technique is said to have construct validity if it can be demonstrated that it measure just the ability which it is supposed to be measured. Two raters have given the scores on the students' argumentative essays based on the criteria assigned. The criteria assigned both in table of specification and scoring guide describe about the abilities of students intended to measure in their essay and this figures portray scores of the students' argumentative essays. Therefore, he reasons that argumentative essay writing test that he has been developed and tried out is to have high construct validity.

E. Analysis of Test Reliability

To analyze reliability of the argumentative essay test, test developer applies inter-rater reliability. Reference [2] states that inter-rater reliability is to estimate reliability level of two sets of scores obtained from two scorers who respectively score the same test-takers. The level of test reliability can only be expressed in the form of quantitative coefficient correlation ranging between +1.00 and -1.00 which is obtained from applying one of the formulas to calculate the level of coefficient correlation. The formula applied in this TDP project is coefficient correlation of Pearson product-moment as follows:

$$r - xy = \frac{\sum(X - \bar{X})(Y - \bar{Y}))}{N\, S_x\, S_y}$$

With this formula, we can calculate the scores of students' argumentative essays scored by two raters—rater1 labeled (X) and rater2 labeled (Y) as in table 5 below. The implementation of the formula can be seen at the bottom of the table in which it elicits Pearson product-moment coefficient correlation value r = 0.82. This value indicates that the test reliability level of argumentative essay test which is developed by test developer can be categorized as in "high level of coefficient correlation." Reference [2] states that inter-rater reliability is to estimate reliability level of two sets of scores obtained from two scorers who respectively score the same test-takers. The level of test reliability can only be expressed in the form of quantitative correlation coefficient ranging between +1.00 and -1.00 which is obtained from applying one of the formulas to calculate the level of coefficient correlation.

The discussion part of the TDP is mainly concerning with inter-rater reliability and validity of the developed test. According to [5], inter-rater reliability occurs when two or more scorers yield inconsistent scores of the same test, possibly for lack of attention to scoring criteria, inexperience, inattention or even preconceived biases. Reference [17] narrated that the rater is also human and a host of contaminating factors affect his rating. On a more technical level, ratings are distorted by the: (1) "halo effect," where the rater evaluates or reacts to each item in the direction of the general impression of the teacher, (2) "error of leniency," a tendency of the rater to rate low or high, no matter the reason, (3) "error of central tendency," whereby the rater reluctantly offers extreme judgments about others (teachers), and (4) "constant error," whereby the rater tends to rate others in the opposite direction to his attitudes and behaviors.

In her study of native and nonnative-speaking EFL teacher's evaluation of Chinese students' English writing, [18] reported that English background teachers attended more positively in their criteria to the content and language, whereas the Chinese teachers attended more negative to the organization and length of the essay. The Chinese teachers were also more concerned with content and organization with their first criteria, where English-background teacher focused more on language in their criteria.

Reference [19] reported from his study of "rater bias in EFL writing assessment" revealed that several recurring bias patterns among subgroups. In rater-category bias interactions, if content and/or organization were rate severely, then language use and/or mechanics were rated leniently, and vise verse. In rater-writer bias interaction, there tended to be more severe or lenient bias towards higher ability writers than lower ability writers. Some writer also rated higher ability writers more severely and lower ability more leniently than expected. Consistent with what [17] stated that raters are not always motivated or honest; moreover, because raters are human, they offer imperfect judgments; they remain susceptible to selective perceptions, memory, and lack of sensitivity to importance or significance. He as well narrated that other factors affecting raters include: (1) sex, (2) race (ethnicity), (3) age, (4) intelligence, (5) understanding of directions, (6) understanding of purpose, (7) sufficient time to complete the ratings, (8) possession of traits measured, and (9) different criteria raters employ for assessing the same trait or behavior.

The different between authentic writing contexts and the writing test situation bring into question the validity of a single writing sample as a predictor of future writing performance in authentic assessment. Breldan in [7] suggests that limited sampling causes greater errors than unreliability problems due to reader's disagreement or inconsistency; "When only one writing sample has been scored, it is not possible to estimate accuracy anything but reading reliability." The developed test possesses this idea in stating that a single writing sample is essentially a one-item test.

Research has shown that different types of wiring topics (prompts) make different cognitive demands on writers and may elicit different type of responses, which may not be assigned equivalent scores. There is no consensus on what constitute a "difficult" or "easy" writing topic in terms of either rhetorical mode of discourse or of content or subject matter. The relationship between these prompt variables are unclear and may in fact change in evaluations of native versus nonnative writing persons with high versus low language proficiency, and secondary school students versus graduate students. Certain rhetorical modes and certain subject or content areas may be more or less difficult to write on in an impromptu situation [7].

Based on the analyses of students' scores on their argumentative writing essays, it reveals that the developed test format have high both validity and reliability. Reference [2] addresses validity as the suitability of test's result with interpretation toward the test as evaluation instrument, however, in a simpler and more practical use, validity is associated with the appropriateness of test as measurement tool with the skills or target intended to be measured. This notion is consistent with what [5]argues that validity is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in term of purpose of the assessment. Thus a valid test of writing is to measure writing ability with some consideration of comprehensibility, rhetorical discourse elements, and organization of ideas, among other factors. Reference [9] briefly states that validity of a test is the extent to which it measures what it is supposed to measure and nothing else. He then elucidates that every test, whether it be a short, informal classroom test or a public examination, should be as valid as the constructor can make it. The test must aim to provide a true measure of the particular skill which it is intended to measure: to the extent that it measure external knowledge or other skills at the same time.

In the mean time, the validity of writing test according to [8] can best addressed in term of construct validity, content representativeness (or validity) and curricular validity. He further argues that construct validity can best be guaranteed by an analysis of the general features of writing situations and a resulting defensible specification of the domain of writing task. This is a functional approach to construct validity and it was used in the IAE International Study of Composition. This idea in line with what [2] states that validity can be identified and verified via observing its suitability with the content (content validity), appropriateness with criteria (criteria validity) and fittingness with construct—in term of concept, approaches and philosophical bases (construct validity).

Reference [5] is questing on "How is the validity of a test established?" He then responses his own question describing that there is no final, absolute measure of validity, but several different kinds of evidence may be invoked in support. In some cases according to him, it may be appropriate to examine the extent to which a test call for performance that matches that of the course or unit of study being tested. In other cases, it may be concerned with how well a test determines whether or not students have reached an established set of goals or level of competence. Statistical correlation with other related but independent measure is another widely accepted form of evidence. Other concerns about a test's validity may focus on the consequences—beyond measuring the criteria themselves—of a test, or even on the test-taker's perception of validity. Reference [9] is further arguing that the test situation or the technique used is always an important factor in determining the overall validity of any test. Although an ideal test situation will by no means guarantee validity, a poor test situation will certainly detract from it.

Reliable test is consistent and dependable [5]. A good test according to [2] should have reliable characteristics such as constant, consistent, and accountable to elicit what is being measured. That is, a test that can generate consistent scores which have no extreme deviation. While, [9] argues that reliability is a necessary characteristic of any good test; for it to be valid at all, attest must first be reliable as a measuring instrument. If the test is administered to the same candidate on different occasions (with no language practice work taking place between these occasions), then, to the extent that it produ,ces differing results, it is not reliable.

Most authors of language testing and evaluation agree that there are some factors can affect the reliability of a language test such as: (1) the extent of the sample materials selected for testing; (2) the administration of the test; (3) test instruction; (4) personal factors such as motivation and illness and (4) scoring the test [2, 9, 10]. According to [9], scoring the test is one of the most important factors affecting reliability. Objective tests overcome this problem of marker reliability, but subjective test are still faced with it; hence the important of the work carried out in the field of the multiple-marking of compositions.

One method of measuring reliability of a test, according to [9], is to re-administer the same test after a lapse of time. It is assumed that all candidates have been treated in the same way in the interval – that they have either all been taught or that none of them have. Another means of estimating the reliability of a test is by administering parallel forms of the test to the same group. This assumes that two similar version of a particular test can be constructed: such as test must be identical in nature of their sampling, difficulty, length rubrics, etc. only after a full statistical analysis of the tests and all items contained in them can the tests safely regarded as parallel. If the correlation between the two tests is high (i.e. if the results derived from the two tests correspond closely each other), then the test can be termed reliable.

The split-half method is yet another means of measuring test reliability. This method estimates a different kind of reliability from that of estimated by test/retest procedures. The split-half method based on the principle that, if an accurate measuring instruments were broken into two equal parts, the measurement obtained with one part would correspond exactly to those of obtained with other. The test is divided into two and the corresponding scores obtained, the extent to which

they correlate with each other governing the reliability of the test as a whole. One procedure widely used is to ascertain the correlation between the scores on the odd numbered items and those on the even numbered items [2, 5, 6, 9, 10].

IV. CONCLUSION

To conclude, underlying on the test reliability analysis, the test developer can now argue that the argumentative essay writing test which he has developed and tried out to have reliability quality. In this case, he can use this test in the practical use to test students' skill in writng argumentative essay. Others can also develop the same test in different genre of writing essay test by modifying components of the test such as; criteria assigned; table of specification; and scoring guide, of course which is based on the general and specific objectives of the course and components of measuring test-takers' ability to write argumentative essay.

REFERENCES

[1] D. Resnick and L. Resnick, "Performance assessment and the multiple functions of educational measurement," Implementing performance assessment: Promises, problems, and challenges, pp. 23-38, 1996.

[2] S. Djiwandono, "Tes bahasa pegangan bagi pengajar bahasa," Jakarta: PT Indeks, 2008.

[3] M. Torrance, G. V. Thomas, and E. J. Robinson, "Individual differences in undergraduate essay-writing strategies: A longitudinal study," Higher Education, vol. 39, pp. 181-200, 2000.

[4] S. K. Kitao and K. Kitao, "Testing Writing," 1996.

[5] H. D. Brown, Language assessment: Principles and classroom practices: Allyn & Bacon, 2004.

[6] A. D. Cohen, "Assessing language ability in the classroom," 1994.

[7] D. Douglas and C. Chapelle, A New Decade of Language Testing Research: Selected Papers from the Annual Language Testing Research Colloquium (12th, San Francisco, California, March 1990): ERIC, 1993.

[8] S. Takala, "Testing Writing Ability: A Review," 1986.

[9] J. B. Heaton, Writing English Language Tests (New Edition). NewYork: Longman Publishing Group, 1989.

[10] J. D. Brown, Testing in Language Programs: A Comprehensive Guide to English Language Assessement: McGraw-Hill College, 2005.

[11] J. Hoetker, "Essay examination topics and students' writing," College Composition and Communication, vol. 33, pp. 377-392, 1982.

[12] G. Fulcher, Testing second language speaking: Pearson Education, 2003.

[13] Panduan Akademik Fakultas Keguruan dan Ilmu Pendidikan 2016, 2016.

[14] R. L. Smalley, M. K. Ruetten, and J. Kozyrev, Refining composition skills: Rhetoric and grammar: Heinle & Heinle Boston, MA, 2001.

[15] A. Oshima and A. Hogue, Introduction to Academic Writing. New York: Addison-Wesley Publishing Company, Inc, 1988.

[16] E. Hatch and H. Farhady, "Research design and statistics for applied linguistics," 1982.

[17] A. C. Ornstein, "Can we define a good teacher?," Peabody Journal of Education, vol. 53, pp. 201-207, 1976.

[18] J. Lee, "Language Testing as a Technique to Enhance EFL Teaching Effects on Vocabulary Acquisition at the Intermediate Level," Faculty of Education. Philadelphia, The Pensylvania State University, 2000.

[19] E. Schaefer, "Rater bias patterns in an EFL writing assessment," Language Testing, vol. 25, pp. 465-493, 2008.