Advances in Social Science, Education and Humanities Research, volume 254
Eleventh Conference on Applied Linguistics (CONAPLIN 2018)

Validity and Reliability Study on Teacher-Made Assessment for English Mid-Term Examination

Naelul Rohmah*
Universitas Pendidikan Indonesia
Bandung, Indonesia
*aynyla@upi.edu

***Abstract*—This study is primarily aimed to study the validity and reliability of teacher-made assessment for English mid-term examination used in evaluating students' learning experience: to what extent the test quality is in term of content validity, reliability, index of difficulty, index of discrimination, and effectiveness of distractors. Qualitative and quantitative data were collected, analyzed, and interpreted by employing a descriptive and exploratory study on teacher-made English mid-term test as well as 20 students' answer sheets as the source of data. The source of data was collected from participants in one private Islamic Senior High School in Kuningan, West Java. Document analysis was conducted to pull out the findings of the study. The test items were observed through its representation in syllabus to reveal its content validity. Then, the students' answers were analyzed using Anates software version 4 to determine the reliability, index of difficulty, index of discrimination, and effectiveness of distractors of the test. The findings of the study show that content validity and reliability were considered good. However, some items were found to be either too easy or too difficult in terms of its index of difficulty. The test also had 50% poor discrimination index and 40% ineffective distractors. Therefore, it is suggested for the teacher to consider a further revision on the English mid-term examination they made for improving of its quality.**

Keywords—summative assessment; test validity; test reliability; item analysis

I. INTRODUCTION

The shift of education system from school-based curriculum to competency-based curriculum (known as curriculum 2013) has made the teaching learning system at schools changing its policies. The assessment technique that the teachers use is one specific instance that was influenced by the changing curriculum. The so-called authentic assessment was introduced as the assessment technique applied in Curriculum 2013. Authenticity and meaningfulness are emphasized during the process of assessment. It measures students‘ learning progress including the three broad aspects: attitudes, knowledge, and skills [1]. Wraggs stated that primary aspects of assessment are knowledge and understanding (related to factual information, concepts, name labels, ideas, theories, applications, connections, etc.); skills (related to techniques, specific competence in particular fields, ability to link knowledge, etc.); attitudes and values (related learning, behaviors, beliefs, subject knowledge, people, and society); and behavior (related to social relationships, personal characteristics, etc.) [2,3]. In curriculum 2013 in line with scientific approach, the teachers are demanded to evaluate each indiviudal student' progress in learning.

Some problems appear when many teachers felt that it is burdensome to fill in many forms of assessment because there are many other responsibilities which the teachers should fulfill related to the teachers‘ professional development [4], involving the administration affairs (syllabi, lesson plans, meeting progress reports, etc) and the pre-teaching preparation (preparing the materials and media). As a result, the teachers admitted that they found it difficult to focus on making a comprehensive test for students, as for instance, when they were assigned to develop test items for English mid-term examination.

The aforementioned issue is supported by Norris statements that language teachers are indeed challenged with the responsibility of developing language tests [5]. However, determining which testing types are the most appropriate for a particular language education context may be daunting [6], especially during the time when the increasing variety of instruments, procedures, and practices are available for language testing.

Furthermore, the assessments are mostly conducted at the end of a semester. Meanwhile, the school system does not concern on the assessment for the mid-semester period [4]. A mid-semester evaluation is considered particularly essential for allowing the teachers to influence the students whom they are currently teaching, while the end-of-term assessments only affect the future semester. Through mid-term assessment, teachers are able to display their interests and concerns on what and how students are learning and their feedbacks to the teaching practice.

Another issue was raised when students complained about how too often times the test results are not so reliable for students obtaining scores that are either higher or lower than they ought to be [7]. Students felt that assessment in some instances is completely or partially unrelated to the contents that they have learned in the class. In other words, there is an inconvenient mismatch between the content learned in class and the material assessed at the end of unit test [8]. They further stated that this may lead to the discredibility of

assessment development based on which teachers make valid judgments about students' progress.

In addition, Fiktorius pointed the issue of the public assumption that examination in Indonesia has been lacking some authenticity [9]; given that authenticity is defined as the extent of relation between the characteristics of the test tasks and real world tasks [10-12]. Thus, the validity of teacher-made assessment remains debatable in the educational assessment process.

Some studies have been carried out in analyzing the validity and reliability of teacher-made assessments. Odiemo's study has found out that the experienced teachers who have gone through trainings on test development and analysis tended to design tests with higher validity and reliability than their counterparts without such training [13]. Moreover, Lei et al study has demonstrated that eventhough the teacher-made assessment was valid in terms of content validity, most of the teachers could not refer to the table of specification while building instruments for assessment due to their low understanding and awareness regarding the importance of table of specification as it enables teachers to see whether or not the assessed objectives actually reflect the ability which the subject is designed to cater [14,15].

Based on these arguments, the questions are raised to the surface, "in terms of validity and reliability, are current teacher-made tests considered having a good quality?" To discover the answer, an analysis of validation should be carried out to impose quality assurance [9,10]. In order to find out the advisability and the quality of the test, analysis of the teacher-made English mid-term examination quality is conducted in terms of investigating its face validity, content validity, reliability, index of difficulty, index of discrimination, and the effectiveness of distractors.

II. METHOD

Two approaches were employed in this research. Qualitative evidence was gathered to assess the content validity, and quantitative data were obtained to reveal the reliability, index of difficulty, index of discrimination, and effectiveness of distractors of the test. This research was descriptive and exploratory in nature. Teacher-made English mid-term test as well as 20 answer sheets of 11th grade Senior High School students were collected as the source of data. Two techniques were pulled out to collect the data, namely a table specification checklist and document analysis. The checklist was one of the instruments used to analyze the test validity. In this case, a test specification was used to analyze content validity and construct the validity of the teacher made test. Meanwhile, the document analysis in this study used the syllabus, the test items, and the students' answer sheets.

To analyze the content validity, the test items were compared to the demands of the syllabus content including the standard competency and basic competency. This would be the proof whether the test items were appropriate to what the students had learned in their classrooms. Reliability, discrimination index, difficulty index, and the effectiveness of distractors were analyzed using Anates Version 4. The Anates software was used to analyze the raw data to get a result. The results from the Anates version 4 analyses were then compared to each criterion via: for reliability, the discrimination index, the difficulty index, and the effectiveness of the distractors.

III. FINDINGS AND DISCUSSION

First, it is concluded that the test has fulfilled the criteria of having a good content validity. The total number of test items is 45 test items with details 40 multiple choices and five essay items. The total main topics of lesson in the curriculum 2013 syllabus of Senior High School 11th grade students are 12 main topics. These are topics in chronological order: expressions of recommendation and offers, expression of asking for and giving opinions, expression of hopes and prayers, formal invitation letter, personal letter, procedural text, passive voice, conditional sentences, factual report, analytical exposition text, biographical text, song analysis. Two items (no. 7 and 12) covered the first topic, expression of recommendation and offer. Two items (8, 10) covered the second topic, expression of asking for and giving opinions. One item covered the third topic (11), expression of hopes and prayers. Two items (13, 14) covered the fourth topic, formal invitation letter. Two items (13, 14) covered the fourth topic, formal invitation letter. Four items (21, 22, 23, and 24) covered the fifth topic, personal letter. Seven items (1, 2, 3, 4, 5, 6, and 43) covered the sixth topic, procedural text. Two items (29, 30) covered the seventh topic, passive voice. Three items (31, 34, and 42) covered the eighth topic, conditional sentences. Three items (37, 38, 39, and 40) covered the ninth topic, factual report. Three items (15, 16, 17, and 18) covered the tenth topic, analytical exposition text. One item (41) covered the eleventh topic, biographical text. Three items (19, 20, 44, and 45) covered the twelfth topic, songs analysis. Thus, 39 test items successfully covered the main topics in the syllabus.

TABLE I. THE COVERAGE OF TEST ITEMS ON MAIN TOPICS IN SYLLABUS

No	Topics Covered in the Syllabus	Items Number	Number of Items In Total	Percentage
1	Expressions of Recommendation and Offers	7, 12	2	4.4%
2	Expression of Asking for and Giving Opinions	8,10	2	4.4%
3	Expression of Hopes and Prayers	11	1	2.2%
4	Formal Invitation Letter	13, 14	2	4.4%
5	Personal Letter	21, 22, 23, 24	4	8.8%
6	Procedural Text	1, 2, 3, 4, 5, 6, 43	7	15.5%
7	Passive Voice	29, 30	2	4.4%
8	Conditional Sentences	31, 34, 42	3	6.6%
9	Factual Report	37, 38, 39, 40	4	8.8%
10	Analytical Exposition Text	15, 16, 17, 18	4	8.8%
11	Biographical Text	41	1	2.2%
12	Song Analysis	19, 20, 44, 45	4	8.8%

As for the coverage of test items on the language elements in syllabus, three items did not cover the main topics in syllabus and its basic competencies. However, the test items were still included in the specific materials in syllabus, which is language elements. One test item (32) covered the question about past continuous tense. Two test items (33, 35) covered the question about simple present tense. Both topics were not specifically stated in syllabus as the basic competencies, but still learned as part of certain main topics. Thus these three items were admitted to the group of items who covered the contents in the syllabus.

TABLE II. THE COVERAGE OF TEST ITEMS ON THE LANGUAGE ELEMENTS IN SYLLABUS

No	Language Elements in Syllabus	Items Number	Number of Items in Total	Percentage
1	Past Continuous Tense	32	1	2.2%
2	Simple Present Tense	33, 35	2	2.2%

For the coverage of test items on topics not in the syllabus, six items failed to cover any content in the syllabus. One item (9) covered the topic of expression apology. Five items (25, 26, 27, 28, and 36) covered the topic of recount text. Both topics did not come under any content in the syllabus. In other words, students did not learn about expressing apology nor recount text during their learning process in 11th grade. Thus, these six items have been out of the criteria of having a good content validity.

TABLE III. THE COVERAGE OF TEST ITEMS ON TOPICS NOT IN THE SYLLABUS

No	Other Topics Not In Syllabus	Items Number	Number Of Items In Total	Percentage
1	Expression of Apology	9	1	2.2%
2	Recount Text	25, 26, 27, 28, 36	5	11.11%

It can be concluded that 39 items covered the contents written in syllabus. All 12 main topics are also encompassed. Whereas, six items did not cover any content in the syllabus. Thus, 87% of test items covered the contents in syllabus, and 13% of test items are irrelevant.

A test has a good content validity if it covers all the contents as stated in the curriculum [16]. If the percentage of representation of test items in content of the curriculum is 50% or more, the test has high content validity [6]. Therefore, the test items can be considered as having a good content validity for the 87% of representation of test items in content of curriculum 2013 syllabus.

Second, as for the reliability of the test, the index of reliability was shown to be 0.79. The reliability is higher than 0.70, then the test is reliable [17]. On the contrary, if the reliability id lower than 0.70, then the test is considered unreliable. Therefore, the test is reliable.

Third, the index of difficulty is shown in three levels: too easy, acceptable, and too difficult. Thorndike and Hagen asserted that a proportion of correct answers less than 0.30 is classified too difficult [18]. While a proportion of correct answers exceed 0.70 is labelled too easy. In other words, any given test items that have the difficulty indexes ranging from 0.30 to 0.70 are considered to be good items.

As the result, six items were labelled too easy, 24 items were acceptable, and 10 items were classified too difficult. Thus, 24 items can be considered to be good items.

Fourth, according to Sudjiono, the index of discrimination that fall below 0.20 is considered poor [9]. Those lying from 0.20 to 0.40 are labelled satisfactory, the classification of good is addressed to those between 0.41 and 0.70. The classification of excellent is labelled to those above 0.71. As the result, eight items were labelled poor, 14 items were satisfactory, six items were good, and 10 items were excellent.

Lastly, the aim of distractors to appear as a plausible option for those students who have not achieved the objective being measured [19]. Hence, a distractor is considered effective if it successfully distracts the students whose performance in the test were below expectation. 40 multiple choice items provide 200 options by each item has five options (a, b, c, d, and e). There are only 40 corrections. Therefore, 160 distractors are available. The result shows that 66 distractors did not work well as they were implausible options. While 94 distractors have worked well as they are plausible options for students who have not achieved the intended objectives.

The findings of study have shown a satisfactory result by proving that the mid-term assessment made by the teacher has successfully reached the level of good validity and reliability. The result supported the previous study by Odiemo, on statement that experienced teacher who went through trainings on test development and analysis has a better test design with higher validity and reliability than their colleagues with less training experiences [13]. In this study, the teacher is admitted to have a five-year English teaching experience and specifically three-year experience of teaching English in Senior High School. Furthermore, despite many teachers complained that designing mid-term assessment just simply adds up to their burdensome abundance of many other responsibilities involving the administration affairs and pre-teaching preparation [4], the findings of this study may have just testified that experienced teacher is still able to manage her responsibilites well by designing a valid and reliable mid-term assessment.

IV. CONCLUSION AND RECOMMENDATION

Based on the results of the validity and reliability study on the teacher-made English mid-term examination, it can be concluded that, first of all, the content validity of the test was good since the percentage of syllabus coverage is 87%. Second, the test had a high degree of reliability. Third, 60% of the test were acceptable in terms of its index of difficulty. Fourth, the discrimination index was equal. 20 test items were ranging from poor to satisfactory, and the other half were ranging from good to excellent. Finally, the effectiveness of distractors was 60% plausible.

In can be concluded that the quality of English mid-term examination made by the teacher was good in terms of its content validity and reliability. Nevertheless, it is suggested for

teacher, as the test developer, to consider revising items as some items were found in the category of 'too easy' and 'too difficult', 50% poor discrimination index, and 40% implausible distractors.

REFERENCES

[1] The Ministry of Education and Culture. Materi pelatihan guru implementasi kurikulum. (The materials for teacher training of curriculum implementation). 2013: SMP-Bahasa Inggris. Jakarta: The Ministry of Education and Culture. 2013.

[2] E.C. Wraggs, Assessment and learning in the primary school (successful teaching series). New York: Routledge. 2001.

[3] M. Briggs, A. Woodfield, C. Martin, and P. Swatton Assessment for learning and teaching in primary schools second edition. Exeter: Learning Matters Ltd. 2009.

[4] I. Fathin, and R Martanti, Holistic english mid-term assessment for junior high school. Indonesian Journal of Language Studies, vol. 1(1), pp. 57-69, 2015.

[5] J. M. Norris, Purposeful language assessment: selecting the right alternative test. English Teaching Forum, vol. 38(1),41– 45. 2012.

[6] N. Kholilah, The quality of english language testing implemented in kbri school, sekolah indonesia kuala lumpur, Malaysia. IJET, vol. 5 (1), pp. 150-172. 2016.

[7] D. A. Frisbie, Reliability of scores from teacher-made test. Article of NCME Instructional Module, University of Lowa. 1988.

[8] H. Fives, and N. DiDonato-Barnes, Classroom test construction: the power of a table of specifications. Practical Assessment, Research and Evaluation, vol. 18 (3), pp. 1-7. 2013.

[9] T. Fiktorius, A validation study on national english of junior high school in indonesia. Master degree thesis, Tanjungpura University, Pontianak. 2014.

[10] L. F. Bachman, Building and supporting a case for test use. Language Assessment Quarterly, vol. 2(1), pp. 1-34. 2005.

[11] L. F. Bachman, A.S. Palmer, Language testing in practice. Oxford: Oxford University Press. 1996.

[12] H.D. Brown, Language assessment: principles and classroom practices. California: Longman. 2003

[13] L. Odiemo, Validity and reliability of teacher-made tests: case study of year 11 physics in nyahururu district of kenya. African Educational Research Journal, vol. 2(2), pp. 61-71. 2014.

[14] M.I. Lei, M.B. Musah, S.H. Al-Hudawi, L.M. Tahir, L.M. Kamil, Validity of teacher-made assessment: a table of specification approach. Asian Social Science, vol. 11(5), pp. 193-200. 2015.

[15] S. Wolming, and C. Wikstrom, The concept of validity in theory and practice. Assessment in Education: Principles, Policy and Practice, vol. 17(2), pp. 117-132. 2010.

[16] J. B. Heaton, Writing english language test. New York: Longman. 1988.

[17] T. Braun, W. Glänzel, and A. Schubert, A Hirsch-type index for journals. Scientometrics, vol. 69(1), pp. 169-173. 2006.

[18] A. Sudjiono, Pengantar Evaluasi Pendidikan. Jakarta: Raja Grafindo Persada, 2008.

[19] S. J. Burton, R. R. Sudweeks, P. F. Merril, and B. Wood, How to prepare better multiple choice tests: guidelines for university faculty. Provo: Brigham Young University Press. 1991.