Validity and Reliability of Semester Tests Made by Teachers: An Evaluation Study of English Learning

M. Aries Taufiq[1], Rahmi Eka Putri[2], Hasri Fendi[3], Syafrilianto[4], Dian Anggraini[5], Vivi Indriyani[6]

[1,2,6]Faculty of Language and Art Universitas Negeri Padang, Padang, Indonesia, [3]Faculty of Abad and Humanities, Universitas Islam Negeri Imam Bonjol, Padang, Indonesia, [4]Faculty of Teaching and Education, IAIN Padang Sidempuan, Padang Sidempuan, Indonesia, [5]Faculty of Economic, Universitas Putra Indonesia YPTK, Padang, Indonesia

[*]ariespertama@gmail.com

Abstract. This research was conducted to analyze and evaluate the quality of semester tests made by grade VIII teachers of junior high school (SMP) at Kec. Tambang for English subject. Thus, it was classified into evaluation studies. The findings revealed that the tests were categorized as valid and had the highest reliability index (0.71). Among the four tests, the semester test used at SMPN 4 got the best quality compared to tests used at other schools. In conclusion, the quality of the semester tests designed by English teachers in class VIII at SMP Negeri Kec. Tambang varied in terms of validity and reliability.

Keywords: evaluation; summative assessment; validity; reliability; teacher-made test

1 Introduction

Evaluation is a superordinate term for measurement and testing [1]. This is a very important feature in determining teacher decisions before education, during education and after education and ensuring control of educational activities [2]. Evaluation relates to the procedure used to determine whether students meet the specified criteria or not [3]. It is used to measure personal success or define student learning [4]. Evaluation is one of the determinants of how successful a curriculum program is in achieving its goals [5]. In this context, quality education ensures quality learning and quality education is only possible through quality evaluation [6].

One of evaluation tools that is commonly used in education is tests. They are designed to assess the quality, abilities, skills or knowledge of a sample against a given standard, which can usually be considered acceptable or not [3]. Tests are a measuring method for someone's ability, knowledge, or performance in a certain domain. Teachers can reveal the students' understanding towards the material through the tests, they can also evaluate the effectiveness of their performance in teaching by giving a test to the students [7]. Apart from being a measure, tests are also a stick to direct students who have insufficient internal motivation and motivation [8].

One form of evaluation commonly used at schools is summative assessment. Summative assessment scores on international assessments often determine the quality of education for a country [9]. Summative assessments relate to evaluations of completed programs [3]. This can

BICED 2020, September 14-15, Bukittinggi, Indonesia

DOI 10.4108/eai.14-9-2020.2305675

be conducted after finishing the unit. Any type of study can be based on the teacher's summative observations of students or test results formalizing their achievements [10]. This assessment aims to record or report student achievements [2]. This is a kind of assessment that summarizes all pieces of evidence up to a point [11], aiming to look back and record how students have achieved the goal [12]. Simply put, this is a type of evaluation that summarizes the strengths and weaknesses of a program [3].

Summative assessments are based on predetermined criteria or standards that used to produce numerical data in the form of test scores [13]. Teachers tend to use test items in class summative tests that focus on remembering and memorizing [14]. Summative assessments are usually given as part of the "test time" specified in an environment designed to eliminate distractions and interactions with other students. This "test time" is intended to be different from the teaching time [15]. In language learning, this summative assessment focuses on mastering linguistic accuracy, and emphasizes linguistic competence rather than communicative competence [16].

The use of standardized tests in education has increased greatly over the past 50 years [17]. The national or state testing systems has been seen as an important strategy for improving the quality of education [18]. This also happens in Indonesia. The semester test is given to students as an instrument to assess their achievement after learning for a semester. As a summative test, the semester test provides teachers evidence about the students' achievement related to the learning goals stated in the curriculum. Currently, the semester tests are designed by a group of teachers (MGMP) or selected teachers in a certain area. Such tests are usually referred to as standard tests because they are designed by qualified teachers and are selected according to their ability to design tests.

This also happened in SMP Negeri Kec. Tambang Kab. Kampar, Riau. The semester tests were usually designed by certain teachers chosen by the MKKS (Principal's Work Meeting). The teacher was chosen based on his/ her competence and experience in designing tests. However, starting from the 2015/2016 academic year, the semester tests, including English, in most of the State Junior High Schools (SMP Negeri) at Kec. Tambang, Kab. Kampar were designed by teachers from each school; this was known as teacher-made tests. Each school was given the responsibility to design their own tests to assess student achievement at the end of the semester. This implies that each school would have a different version of semester tests. In this case, teacher-made tests play a central role in evaluating students' learning achievement at junior high schools in Kec. Tambang, Kab. Kampar, Riau.

Generally, teachers are considered as the key actors at schools, contributing and shaping the students' development and learning [19]. It is a common belief that good teachers are good test developers [20]. However, in preparing a test, some of the following obstacles are found by the teachers. First, there is a tendency for teachers to use test items in class summative tests that focus on memory and memorization; therefore, it influenced the teacher's disposition towards summative assessment as well [21]. Second, under increasing pressure to improve the students' scores, teachers are more likely to use shortcuts or limit the instruction to test certain contents and activities [19]. Third, teachers rarely discuss or share their practices with colleagues at the same school. Teachers not only are not aware of the practice of their coworkers, but also do not trust the evaluation results obtained from their peers [20]. Fourth, scores obtained from summative assessments tend to predict invalid and unreliable scores about students [13].

Determining the test objectives and choosing the most suitable type of test must be done by the teacher before giving the test to the students in order to conduct an effective test. However, it is not enough to have only an effective test. In this case, educators must first

consider certain principles, such as legitimacy and reliability. These two aspects are the most important among other principles [22]. Tests can be considered valid if they measure what is intended to be measured [19]. Meanwhile, a reliable test means having the consistency of the scores obtained from the test (Scheerens, Glas, & Thomas, 2005). Reliability is a measure of how stable, reliable and consistent a test is in measuring the same thing every time [23].

Furthermore, validity is a essential criterion that must be considered in developing tests [24]. Validity consists of Face validity, Content validity, Criterion validity, and Construction validity. Face validity means that by just looking at the test, it is possible to conclude that the test is valid [23]. Content validity is estimated through testing the feasibility of the test content through rational analysis by a competent panel of experts [25]. This measurement are accommodated to the measurement specifications in the form of measuring instruments and the type of understanding that is measured [26]. Then, criterion validity is the validity used when test scores can be linked to other standardized criteria [24]. At last, construct validity ensures that the test actually measures the intended attribute and not other foreign attributes [23].

Interpretation of test scores, if valid and reliable, exposes the effectiveness of teaching and the progress of students, guides the learning evaluation process, meets the expectations of community discourse, and facilitates the realization of institutional and national goals [13]. Based on these explanations it is important to assess the validity and reliability of teacher-made tests in order to achieve the expected educational goals. Based on that, this article presents an analysis of semester tests made by garde VIII teachers in SMP Negeri Kec. Tambang, Kab. Kampar, Riau.

2 Methodology

This research was an evaluation research since this research analyzed data orderly about the quality, and effectiveness of a program or product in order to measure the effect of a program against the goals it set out in order to improve future programming. Hence, this research evaluated the semester tests made by grade VIII teachers whether they were fulfilled the criteria for good tests.

The data of this research were obtained from the semester tests made by grade VIII teachers of SMP Negeri at Kec. Tambang as well as the students' answer sheets and transcription of interviews with teachers who designed the tests. Thus, the source of data was grade VIII students of SMP Negeri at Kec. Tambang who did the semester tests designed by the teachers. There were 588 students from four different schools namely SMPN 1, 2, 4 and 6. From those large numbers, 248 students were chosen as the sample of the research. In addition, the source of data was also English teachers who designed the semester tests for academic year 2015/ 2016. There were 3 teachers chosen for this research.

A checklist format was used to gather the data and to analyze the validity of the tests. The data were gathered by asking the necessary documents such as semester tests and students' answer sheets to the authority. There were four semester tests taken as the source of data which were used by SMPN 1, 2, 4 and 6. Then, students' answers were calculated to get some statistical information. Moreover, to check the trustworthiness of the data, the result of data analysis was discussed with some teachers at SMP Negeri Kec. Tambang. This procedure was conducted to make the judgment objective to achieve the data trustworthiness.

After that, the data were analyzed through several steps. First, the test was analyzed in terms of the validity and reliability. The validity was analyzed by checking the items in the tests with basic competences stated in the curriculum while reliability of the tests was counted by using KR-21 formula.

3 Result and Disscussion

The resuts of data analysis revealed that teacher-made tests at SMP Negeri Kec. Tambang did not fulfill the criteria of validity and reliability. Based on the result, it was found that from the four tests, no test contained a hundred percent valid items. Two tests comprised more than 50% valid items (Test 2 for 62.5% and Test 3 for 66%) while the other two contain valid items below 50% (Test 1 for 41.38% and Test 4 for 34.04%). After that, in term of reliability, the analysis results concluded that generally the tests cover low reliability. To be clearer, the result of data analysis is presented as follow:

Table 1. The Reslt of The Research

Tests	Valid Items (%)	Reliability
1	41.38%	Low (0.32)
2	62.50%	Low (0.41)
3	66.00%	Moderate (0.71)
4	34.04%	Low (0.42)

These findings show that each kind of tests has different quality. These results are discussed as follows:

A. Validity

The discussion of validity was started with the content validity based on the basic competences stated in the curriculum. This research revealed that the teacher-made tests were valid in term of the content since the empirical data indicated that most of basic competences stated in the curriculum were covered in the test items. However, there were several items which were not related to the materials in the syllabus. These unrelated items undermine the test quality since a good achievement test which is valid in the content is supposed to measure the content area in the curriculum. In fact, content validity is the most important type of validity for achievement tests. Therefore, test designers should ensure that the achievement test is valid in the content. In addition, the researchers also checked the item validity and found that none of the tests contains a hundred percent valid items. This was due to several factors. The factors that mostly threaten the test validity were unclear instructions, ambiguous test items and untaught materials included in the test. This finding was supported by Gay and Airasian (2011) asserting that factors that threaten the validity of a test are imprecise test directions, unclear and vague test items, difficult vocabulary, excessively challenging and multifaceted sentence structure, unreliable and biased scoring methods, untaught items included on the tests, failure to follow standardized administration procedures, and cheating [27].

B. Reliability

The results of data analysis showed that the reliability coefficient of Test 3 was Moderate while the reliability coefficient of other tests (Test 1, 2, and 4) was Low. This implies that most of the tests had low reliability. The findings of this research were supported by Arikunto who stated that teacher-made tests usually had low or moderate reliability [28]. This was due to the fact that the teachers had different ability and experience in designing tests. Moreover, the findings of this study showed that these tests could not achieve the criteria of qualified tests. This was supported by the results reliability analysis finding that most of the tests got low reliability. In fact, a good test should have high reliability. This was supported by Ary, Jacobs, and Asghar, stating that the reliability of the achievement test should be in the level of high criteria with the index between $\geq 0.91 - 1$ [29]. The level of reliability criteria of a test becomes evidence that the test is consistent in the results so that the test score can be considered as the real students' achievement. Thus, it can be concluded that semester tests designed by teachers are not reliable since the tests could not provide consistent effects on students' achievement. this implies that the result of the test could be the only consideration for the teachers to evaluate students' achievement.

4 Conclusion

All in all, this research concluded that the semester tests designed by English teachers at SMP Negeri Kec. Tambang could not yet be categorized as effective tests since most of the tests did not accomplish the criteria of good tests, especially their validity and reliability. Furthermore, this research also found that most the test items were required to be revised to make the tests more effective. Hence, the researchers suggested the English teachers to conduct a pilot test before utilizing the test as summative assessment. The teachers could also discuss the content of the test with other teachers and revise, when necessary, items which could bring ambiguous perceptions for the students. The teachers need to ensure the validity and reliability of tests since they are going to be used as an instrument to measure students' achievement. Moreover, the teachers should be involved actively in trainings related to test construction. Hence, it is the principle's responsibility to frequently invite the experts on evaluation or language testing in order to provide the teachers about knowledge required in test constructions.

References

[1] B. K. Lynch, "Rethinking assessment from a critical perspective," *Lang. Test.*, vol. 18, no. 4, pp. 351–375, 2001.

[2] H. S. Yüksel and N. Gündüz, "Formative and summative assessment in higher education: Opinion and Practice of Instruction," *Eur. J. Educ. Stud.*, vol. 3, no. 8, pp. 336–356, 2017.

[3] D. Adom, J. A. Mensah, and D. A. Dake, "Test, measurement, and evaluation : Understanding and use of the concepts in education," *Int. J. Eval. Res. Educ.*, vol. 9, no. 1, pp. 109–119, 2020.

[4] D. Wiliam, C. Lee, C. Harrison, and P. Black, "Teachers developing assessment for learning: Impact on student achievement," *Assess. Educ. Princ. Policy Pract.*, vol.

11, no. 1, pp. 49–65, 2004.
[5] A. M. Coleman, “The dictionary of psychology,” *Appl. Cogn. Psychol.*, vol. 15, no. 3, pp. 349–351, 2001.
[6] Z. Nartgun, “Student Views on the Assessment Practices of Instructors during Instruction,” *Educ. Sci. Theory Pract.*, vol. 9, no. 4, pp. 1807–1818, 2009.
[7] H. D. Brown, *Language assessment: Principles and classroom practice (second edition)*. New York: Pearson Longman, 2010.
[8] D. (W. . Peac (Peachy)i, “Writing evaluation in university English preparatory programs : Two universities of Turkey and Saudi Arabia,” *J. Lang. Linguist. Stud.*, vol. 16, no. 1, pp. 253–264, 2020.
[9] S. Murgatroyd and P. Sahlberg, “The two solitudes of educational policy and the challenge of development,” *J. Learn. Dev.*, vol. 3, no. 3, pp. 9–21, 2016.
[10] R. Ş. Arslan and M. Üçok-Atasoy, “An investigation into EFL teachers’ assessment of young learners of English: Does practice match the policy?,” *Int. Online J. Educ. Teach.*, vol. 7, no. 2, pp. 468–484, 2020.
[11] M. Taras, “Assessment-summative and formative–some theoretical reflections,” *Br. J. Educ. Stud.*, vol. 53, no. 4, pp. 466–478, 2005.
[12] F. Ahmed, S. Ali, and R. A. Shah, “Exploring variation in summativeAssessment : Language teachers ’ knowledge of students ’ formative assessment and its effect on their summative assessment,” *Bull. Educ. Res.*, vol. 41, no. 2, pp. 109–119, 2019.
[13] Z. Ahmad, “Summative assessment, test scores and text quality: A study of cohesion as an unspecified descriptor in the assessment scale,” *Eur. J. Educ. Res.*, vol. 9, no. 2, pp. 523–535, 2020.
[14] C. Klenowski, V.Wyatt-Smith, “The impact of high stakes testing: the Australian story,” *Assess. Educ. Princ. Policy Pract.*, vol. 9, no. 1, pp. 65–79, 2011.
[15] G. Zheng, S. E. Fancsali, S. Ritter, and S. R. Berman, “Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics,” *J. Learn. Anal.*, vol. 6, no. 2, pp. 153–174, 2019.
[16] K. Shaaban, “Assessment of young learners,” *English Teach. Forum*, vol. 43, no. 1, pp. 34–40, 2005.
[17] W. C. Smith, “The global transformation toward testing for accountability,” *Educ. Policy Anal. Arch.*, vol. 22, 2014.
[18] D. Chapman and C. Snyder, “Can high stakes national testing improve instruction: Re-examining conventional wisdom,” *Int. J. Educ. Dev.*, vol. 20, pp. 457–474, 2000.
[19] W. C. Smith and K. Kubacka, “Education policy analysis archives Appraisal Systems,” *Educ. Policy Anal. Arch.*, vol. 25, no. 86, pp. 1–29, 2017.
[20] I. S. Alfallay, “Test specifications and blueprints: Reality and expectations,” *Int. J. Instr.*, vol. 11, no. 1, pp. 195–210, 2018.
[21] E. Bijsterbosch, J. Van Der Schee, W. Kuiper, and T. Béneker, “Geography teachers’ practices regarding summative assessment: A study of pre- vocational education in the Netherlands,” *RIGEO*, vol. 6, no. 2, pp. 118–134, 2016.
[22] H. Öz and T. Özturan, “Computer-based and paper-based testing : Does the test administration mode influence the reliability and validity of achievement tests ?,” *J. Lang. Linguist. Stud.*, vol. 14, no. 1, pp. 67–85, 2018.
[23] Kinyua and Okuya, “Validity and reliability of teacher-made tests: case study of year 11 physics in Nyahururu District of Kenya,” *African Educ. Res. J.*, vol. 2, no. 2, pp. 61–71, 2014.
[24] Suciati, S. Munadi, Sugiman, and W. D. R. Febriyanti, “Design and validation of

mathematical literacy instruments for assessment for learning in Indonesia," *Eur. J. Educ. Res.*, vol. 9, no. 2, pp. 865–875, 2019.
[25] S. Azwar, *Validitas dan Reliabilitas*. Yogyakarta: Pustaka Pelajar, 2007.
[26] D. S. Naga, *Teori skor pada pengukuran mental*. Jakarta: Nagarani Citrayasa, 2012.
[27] L. R. Gay and Airasian, *Educational research: Competencies for analysis and application (10th ed)*. New Jersey: Pearson Education, Inc., 2011.
[28] Arikunto, *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara, 2009.
[29] D. Ary, L. C. Jacobs, and C. K. Sorensen, *Introduction to research in education*, 8th ed. Belmont, USA: Wadsworth, 2010.