Edukasi Jurnal Pendidikan dan Pengajaran
P-ISSN : 2355 - 3669
E-ISSN : 2503 - 2518

Teachers' Perspectives towards Validity of Teacher-Made Test

Hasnan Yasin, Septia Tri Gunawan, Nida Husna, Didin Nuruddin Hidayat
UIN Syarif Hidayatullah Jakarta

Abstract
Some studies that had been conducted showed that teacher-made tests were good and satisfactory. However, the majority of teachers did not validate their tests before administering them to the students. This study was conducted to investigate the perspective of teachers towards their-own-made (teacher-made) tests they had made. The purpose of this study was to know to what extent their agreement regarding their attitudes, quality, and use of the tests. The method used was qualitative descriptive analysis. Five English teachers from Greater Jakarta (Jabodetabek region) participated in this research. The data were gathered through a questionnaire. Their view on the test they have made was analyzed and it was then described. The results showed that (1) the teachers agreed about the appropriateness of the test they administered; (2) the teachers believed that the data quality obtained during research was useful and meaningful, and (3) the teachers used the test to identify and to evaluate their learning objectives, students' learning needs, students' learning difficulties, and school evaluation.

Keywords: English, teachers' perspective, teachers' made test, validity

Manuscript submitted: ***September 24, 2020***
Manuscript revised: ***November 4, 2020***
Accepted for publication: ***December 1, 2020***

Introduction

In Indonesia, anincredibly diverse and multicultural country, English is regarded as one of the most popular foreign. As an assessment approach, standardized tests played a crucial position both in EFL and ESL curriculum evaluation and student evaluation. Brown and Abeywickrama (2010) claimed that an exceptional standardized test was an outcome of practical research and improvement beyond merely acknowledging particular standards or benchmarks. This type of test also entailed systematic procedures for administration and scoring. Most schools around Indonesia employed standardized tests to evaluate students at each level of their educational proficiency. In some cases, particular entities, such as the Board or Ministry of Education and Culture, developed and administered standardized tests. Meanwhile, in other parts with different policies, the tests were administered by the departments within the schools (Akiyama, 2004).

A valid instrument was determined the quality of the test when it was adequately conducted to measure what was deemed to be measured (Muijs, 2011). When an instrument accurately measured any destined variable, it was considered a valid instrument for that particular variable. Jackson (2003) mentioned at least four types of validity: face validity, criterion validity, content validity, or construct validity, as one of the important parts in determining good instruments. Face validity focused on the concept of whether the test seemed valid or not on its facade (Jackson, 2003). Criterion validity was

a notion that would be displayed in the actual study to build it required a good knowledge of theory associating the idea and measuring the relationship between the measurement and the factors related to it. Meanwhile, content validity addressed the content of items, whether it computed the concept being gauged in the study or not. Lastly was the construct validity, which covered the extent of an instrument, so it would precisely measure the theoretical construct composed to figure out the score's amount.

The validity concept could be formulated as to how significantly a test measures what it was meant to be measuring. Valid evaluations produced data that could be used to provide input to educational agreements at various rates, from school advancement to teacher evaluation for individual student earnings and fulfillment. Nevertheless, Caffrey (2009) contended that validity was not an attribute of the test on its own; instead, validity referred to the extent to which specified conclusions drawn from the test results could be perceived as meeting a purpose or situation and important requirements. The validation included collecting facts to justify the use and interpretation of test results based on the principles that the test was intended to assess, defined as buildings. Suppose a test did not measure all the capacity within a concept. In that case, the judgments described from its test results might accurately reflect on the student's knowledge and thus be in place of a validity fulmination.

When the test as an evaluation tool had been proven to have a clear description of the expertise and abilities and aimed to evaluate, an evaluation was considered accurate. It should be part of the validity process that when monitoring for a wide variety of learners, it should also be both compatible with the norm in determining the students' skill and calculable over test settings and scorers (Darling-Hammond, Herman & Pellegrino, 2013). Furthermore, types of data for validity assessment might include: (1) evidence of alignments, as in a statement from a functionally reliable unbiased adjustment study substantiating coordination between the evaluation and its test design, and the criteria of the authority; (2) justification of the validity of using test results for their main objectives, such as a consideration of validity in an authoritative declaration that affirms the aims of the tests, the intended explanations and the use of results; and (3) justification that scores are associated with possible external factors as anticipated, acting as summaries of investigations showing positive relationships with a) external assessments gaging similar constructs, b) student readiness teacher verdicts, or c) test-taker academic attributes.

Evaluation is a daily-based work in the classroom and employed to be a guidance of the teaching-learning process. Evaluation was considered as an instrument or a process used to understand or quantify something in certain situations using certain rules (Arikunto, 2005). Additionally, an evaluation or test might be worth assessing a person's particular aptitudes and skills (Hopkins et al., 1990). Consequently, both evaluation and tests were used as a part of a particular process that educators and examiners' effort in attempting and quantifying students' ability by demonstrating some of the criteria as the sign of the skills being tested (Hedge, 2008).

There were several types of tests that the teachers usually used to see whether or not the learning objectives were achieved. One of them is the achievement test. Achievement test could be categorized into two different types, the standardized evaluation and non-standardized test. A standardized test was when administering the test was prescribed and properly defined (Turkstra et al., 2005). Meanwhile, a non-standardized test was where the process served as an assortment of purposes, such as determining elements in domains where there was no standardized evaluation, describing performance from the context of real-world settings and cognitive requirements and supposed supports (Turkstra et al., 2005).

One that was included as a non-standardized evaluation was the teacher-made evaluation (also referred to as the classroom evaluation). Teacher-made evaluations were developed by topic teachers in schools or universities to rate pupils' achievement in regions covered in education. Usually, it would be limited to be based on a specific topic or group of pupils. While the standardized evaluation was valid, dependable, and contained a table of criteria, the teacher-generated test did not necessarily go through complete sorts of standardization (Okpala et al., 1993). The standardized evaluation was usually made to be used on a far bigger scale compared to teacher-made tests. Therefore, it was exposed to a string of standardization procedures until they were administered on pupils. The teacher-made test used here was included the daily tests, midterm test, and final term test, which had been administered to the students on a smaller scale.

There was plenty of research that had been conducted on the validity of teacher-made tests. Nurhalimah et al. (2019) researched the quality of English teacher-made tests. Her findings showed that most items (80%) in teacher-made tests were in the rate of excellent, good, and satisfactory. This was one of the evidence that teacher-made test quality should not be not taken for granted. In her research, she also gave some comparison to standardized tests. It had been proved that 50% of standardized test items were irrelevant, while teacher-made tests were more superficial. It could be one reason why high scores in schools obtained lower scores in national examinations (Razali & Jannah, 2015). However, most teachers did not validate their test items before administering (Ugwu & Mkpuma, 2019). Despite the urgency and impact of the validity in testing (Friberg, 2010), some teachers still did not consider their test validity. Since no previous studies above explored teachers' notion on a test, the present study attempted to examine the teachers' perspective towards their-own-made (teacher-made) test by formulating several research questions: (a) what are teachers' attitudes towards the appropriateness of the test? (b) what are teachers' perceptions of accuracy in a test? (c) what are the uses of a test for teachers? This inquiry was expected to provide the teachers' agreement regarding their attitudes, quality, and use of the tests.

Methods

Research design

This study used a qualitative method with a descriptive analysis to analyze teachers' perspectives on teacher-made test validity. It was used to describe teachers' perspectives about the teacher-made test and its effect on their teaching. This study's focus was on the perspective or view that was owned by the teachers of the teacher-made test. It was meant to demonstrate the approaches used to determine the validity of the test used by the teachers. This research participants were five English teachers who had been teaching for one to ten years, and they were from Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek) area. The participants were chosen by applying a random sampling technique based on availability, and they had a habit of making their tests.

Data collection and analysis

An adapted questionnaire from Kyriakides (2004) was deployed to see the extent to which the teachers view the validity of the teacher-made test. It was important to remember that the data used was a teacher-made test implemented in the final test. After administering the test, the teacher was asked to fill out a questionnaire. The questionnaire covering the following issues, such as (a) the teacher's attitude towards the suitability of test; (b) teacher attitudes towards the quality of data obtained from tests; and (c) the use of the test by teacher, was adapted from Kyriakides (2004). The questionnaire reliability results were assessed by measuring Cronbach Alpha values relative to the

scale used to assess the teacher's perspective on teacher-made assessments. To measure the teacher responses, Cronbach Alpha is valued for the five scales used in the questionnaire (Cronbach, 1990).

Findings

Five teachers participated in this research by answering the questionnaire related to teacher attitudes regarding the test's appropriateness, teacher attitudes regarding the quality of the data obtained from the test and the test used by teacher.

Teacher attitudes toward the appropriateness of the test

Table 1. Responses of teacher attitudes towards the appropriateness of the test means and standard deviations

Descriptive Statistics			
	N	**Mean**	**Std. Deviation**
the usefulness of data	**5**	**4.40**	**.548**
the evaluative criteria	**5**	**4.40**	**.548**
the scoring	**5**	**4.20**	**.837**
Valid N (listwise)	**5**		

Data from Table 1 was based on the teacher's responses in answering the questionnaires that were related to the suitability of each test activity. The teacher's assessment of the suitability of the testing activity might be based on the suitability of the information collected, the suitability of the topic being assessed and evaluative criteria, and the assessment guidelines' openness. Thus, the teachers were to rate items on a 1 (absolutely disagree) scale to 5 (absolutely agree).

The result showed that the average value of items in line one was high (4.40), where the maximum score was 5, and the standard deviation was relatively low. This showed that, on average, most respondents agreed about using the tests they had made as providing information about their students' literacy skills. Second, the average value of items in line two shows the same value as the previous one (4.40), which shows that evaluative criteria are appropriate. Finally, the average score is lower than the previous two (4.20); therefore, it is still of high value and shows that the teacher considers the assessment guidelines for the tests to be beneficial.

Teacher attitudes towards the quality of the data obtained from the test

Table 2. Percentages of respondents concerning perceptions of accuracy of test result and factors that influenced students' test result, their means, and standard deviations

Descriptive Statistics					
	N	Agree*	Disagree**	Mean	Std. Deviation
Test scores give me some idea of students' literacy	**5**	**60%**	**20%**	**4.00**	**.707**

ability					
The test is indiscriminate and too simple	**5**	**20%**	**80%**	**1.80**	**.447**
Test scores only rank students compared to each other	**5**	**20%**	**80%**	**3.00**	**1.414**
Test scores in each activity reveal the learning needs of each student in specific aspects of literacy assessed by the test	**5**	**100%**	**0%**	**4.40**	**.548**
Student scores are affected by the fact that students are not interested in demonstrating their skills	**5**	**60%**	**40%**	**3.80**	**1.095**
Teacher's knowledge about the individual student is critical to the interpretation or meaning	**5**	**100%**	**0%**	**4.40**	**.548**

given to student's responses to test activities					
Student scores are affected by the context of each test activity, which is familiar only to some groups of students (e.g., middle-class rather than working-class students)	**5**	**40%**	**40%**	**2.60**	**.894**
Student scores are affected by the fact that students are not familiar with the form of activities included in the test	**5**	**40%**	**60%**	**2.80**	**1.643**
Student scores are affected by anxiety	**5**	**40%**	**60%**	**3.00**	**1.414**
Teacher-made test doesn't portray minority language to	**5**	**20%**	**60%**	**2.40**	**1.140**

students accurately	
Valid N (listwise)	**5**

*The respondents either agree or absolutely agree
**The respondents disagree or absolutely disagree

Most respondents (80%) did not think that the test was indiscriminate and too simple. Additionally, they argued that test results were not only useful compared to each other in ranking students but were also useful in giving teachers ideas about their student literacy skills. All respondents (100%) supported the idea that student scores from tests expressed students' learning needs in particular literacy aspects. This assumed that the teacher found the test result was useful for both the formative and summative evaluation functions.

However, more than 40 percent of respondents accepted that grades were affected by the fact that students did not want to show their skills. Furthermore, all respondents believed the knowledge of individuals influenced that test results. About half of the respondents also believed that their familiarity with some students' tests affected the score results. 40% of respondents also supported the idea that student scores were influenced by anxiety. Finally, more than half of respondents believed that this test accurately described minority languages.

The uses of test by teachers

This described how the teachers used the test further. The data of the respondents were shown in the following table.

Table 3. The use of test by the teacher and its means and standard deviations

Descriptive Statistics			
	N	Mean	Std. Deviation
The test helps teachers to identify whether the objectives were achieved	**5**	**4.60**	**.548**
The test helps teachers to identify the learning need of students	**5**	**4.60**	**.548**
The test helps teachers to identify students learning difficulties	**5**	**4.00**	**1.225**
The test is used for summative reasons	**5**	**3.80**	**1.095**
The test is used as sources for evaluating the effectiveness of the school	**5**	**4.60**	**.548**
Valid N (listwise)	**5**		

Table 3 showed that the average test helped the teacher identify whether goals were achieved, students' learning needs, and evaluation of high school effectiveness (4.60) and low standard deviations. This showed that the test or test results were used to do all three things. It also helped teachers to identify student difficulties and was used for summative reasons. This was based on the average value of 4.00 and 3.80, respectively.

Discussion

The information mentioned above could be explained in terms of its impacts on the improvement of teacher-made tests and, in particular, to increase the usefulness of information resulting from tests in decision making and provide appropriate explanations for basic assessment. Moreover, a more general question emerged concerning the significance of examining the instructor's definition as a way of determining the validity test.

First, teachers agreed that this exam offered a depth of information on their students' literacy skills. This idea agrees with Brown (2003), who stated that the test's function is to measure a person's ability, knowledge, and performance. They also often considered evaluative criteria to measure student responses to each related test operation. Furthermore, criteria for evaluation of almost all tasks were found very helpful. This indicated that the teacher seemed to assume that the teacher-made test had to evaluate students for its validity. In addition to that, some research findings showed that most of the teacher-made tests that had been administered to students indicated valid (e.g., Irhamsyah, 2020; Sugianto, 2017). On the other hand, another research found that the teacher-made assessments' validity was low (Minda, 2018). One among other reasons for these varieties is the teachers. As mentioned in a research, the experienced teachers who have gone through training on test development and analysis tended to design tests with higher validity and reliability than their counterparts without such training (Odimo, 2014).

Second, the teacher claimed that the teacher-made test provided information on students' literacy that was to be evaluated. The findings were seen as offering more facts about the validity of the test. The teacher also believed that several factors affect student test scores or results, including student interest, teacher knowledge about individuals, the context of the test, students' familiarity with several groups of students, and anxiety. It is in line with what had been found in several research studies indicating that test scores are affected by many factors (El-Omari, 2016; Farooq et al., 2011; Jurkovic, 2010; Khamkhien, 2010; Shvidko et al., 2015).

Third, it could be claimed that teachers used tests or test scores to identify or evaluate whether goals were achieved, students' learning needs, students' and learning difficulties. According to other research, test scores could be used in evaluating students and teachers themselves as part of a broader form of teachers' evaluation (Baker, 2010; Corcoran, 2010) or even principals of the school (Grissom et al., 2015). This test could also be viewed as a summative reason and a school evaluation for its effectiveness.

Conclusion

To conclude, the gathered data showed that (1) the teachers agreed with the appropriateness of the test they administered; (2) the teachers also believed the quality of data obtained during research was useful and meaningful; and (3) the teachers used the test to identify and evaluate learning objectives, students' learning needs, students' learning difficulties, and school evaluation. Therefore, it was reasonable to assume that validity was a crucial aspect of constructing a test. It is strongly suggested that teachers do some validity test at least randomly if regularly is difficult to conduct. However, one thing that may become a problem, many teachers use the same test several times, and they are barely changing into new ones. Whether or not the validity of their tests is in line with those tests' reliability, it may become a further area to be researched.

References

Akiyama, T. (2004). *Introducing EFL speaking tests into a Japanese senior high school entrance examination*. Melbourne: University of Melbourne.

Arikunto, S. (2005). *Dasar-dasar evaluasi pendidikan*. Indonesia: Bumi Aksara.

Baker, E. L. (2010). *Problems with the use of student test scores to evaluate teachers*. Retrieved from: http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Brown, D., & Abeywickrama, P. (2010). *Language assessment principles and classroom practices* (2nd ed.). White Plain,NY: Pearson Education.

Brown, H. D. (2003). *Language assessment principles and classroom practices*. London:Longman.

Caffrey, E. (2009). *Assessment in elementary and secondary education: A Primer*. English: Congressional Research Service). Retrieved from: https://fas.org/sgp/crs/misc/R40514.pdf.

Corcoran, S. P. (2010). *Can Teachers Be Evaluated By Their Students' Test Scores? Should They Be? The Use Of Value-Added Measures Of Teacher Effectiveness In Policy And Practice.* Texas:Education Policy for Action Series.

Cronbach, L. J. (1990). *Essentials of psychological testing*. New York,NY: Harper & Row.

Darling-Hammond, L. Herman, J., & Pellegrino, J. (2013). *Criteria for high-quality assessment*. Stanford,CA: Stanford Center for Opportunity Policy in Education.

El-Omari, A. H. (2016). Factors affectingsStudents' achievement in English language learning. *Journal of Educational and Social Research*, *6*(2), 9–18.

Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A Case of Secondary School Level. *Journal of Quality and Technology Management*, *7*(2), 1–14.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions?. *Child Language Teaching and Therapy*, *26*(1), 77–92.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3–28.

Hedge, T. (2008). *Teaching and learning in the language classroom*. Oxford ,UK: Oxford University Press.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and Psychological Measurement and Evaluation*. Boston: Prentice Hall.

Irhamsyah, L. H. (2020). The analysis of the teacher-made test for senior high school at State Senior High School 1 Kutacane, Aceh Tenggara. *Jurnal Ilmiah DIDAKTIKA*, *21*(1), 10–20.

Jackson, S. L. (2003). *Research methods and statistics: A critical thinking approach*. Wadsworth : Thomson Wadsworth.

Jurkovic, V. (2010). Language learner strategies and linguistic competence as factors affecting achievement test scores in English for Specific Purposes. *TESOL Journal*, *1*(4), 449–469.

Khamkhien, A. (2010). Factors affecting language learning strategy. *Electronic Journal of Foreign Language Teaching*, *7*(1), 66–85.

Kyriakides, L. (2004). Investigating validity from teachers' perspectives through their engagement in large-scale assessment: The Emergent Literacy Baseline Assessment project. *Assessment in Education: Principles, Policy & Practice*, *11*(2), 143–165.

Minda, M. H. (2018). Content Validity of EFL teacher-made assessment: The case of Communicative English Skills Course at Ambo University. *East African Journal of Social Sciences and Humanities*, *3*(1), 41–62.

Muijs, D. (2011). *Doing Quantitative Research in Education with SPSS*. London,UK: SAGE Publications.

Nurhalimah, N., Fahriany, F., & Dadan, D. (2019). Determining the quality of English teacher-made

test: How excellent is excellent? Indonesia. *Indonesiann EFL Journal: Journal of ELT, Linguistics, and Literature,5*(1), 24-38.

Odimo, L. (2014). Validity and reliability of teacher-made tests: Case study of year 11 physics in nyahururu district of Kenya. *African Educational Research Journal*, *2*(2), 61–71.

Okpala, P. N., Onocha, C. O., & Oyedeji, O. A. (1993). *Measurement and evaluation in education*. Stirling: Horden Publishers.

Razali, K., & Jannah, M. (2015). The comparison between National Final Examination test items and English teacher made-test items of 2010 and 2011. *Al-Talim Journal*, *22*(1), 10–22.

Shvidko, E., Evans, N. W., & Hartshorn, K. J. (2015). Factors affecting language use outside the ESL classroom: Student perspectives. *SYSTEM*, *51*, 11–27.

Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *3*(2), 22–38.

Turkstra, L. S., Coelho, C., & Ylvisaker, M. (2005). The use of standardized tests for individuals with Cognitive-Communication Disorders. *Seminars in Speech and Language*, *26*(4), 215–222.

Ugwu, N. G., & Mkpuma, S. O. (2019). Ensuring quality in education: validity of teacher-made language tests in secondary schools in Ebonyi State. *American Journal of Educational Research*, *7*(7), 518–523.