LINGUISTS

linguist: j. linguist. lang. teach. **5 (1) 2019 29-35**

Linguists: Journal of Linguistics and Language Teaching

http://ejournal.iainbengkulu.ac.id/index.php/linguists

A CRITICAL EVALUATION OF ENGLISH TEST IN GOVERNMENT-AFFILIATED COLLEGE

Monalisa Pasaribu

Institut Teknologi Del Tobasa

monalisa.pasaibu@del.ac.id

Received: June 3rd 2019 Accepted: July 18th 2019 Published: July 2019

Abstract

This study aims at evaluating an English test as a part of entrance test in STAN, a government-affiliated college in Indonesia. STAN EPT has four main sections in the test including Structure and Written Expression, Cloze Question, Error Recognition and Reading Comprehension. Using five criteria by Brown & Abeywickrama (2010), the test is evaluated in terms of its practicality, reliability, validity, washback and authenticity. The result indicated that the test has high practicability with all of questions provided in multiple choices. Some elements of reliability and validity are also covered in this test. Yet, the evaluation also showed that washback and authenticity are not covered in STAN EPT.

Keywords: STAN EPT, English test, test evaluation

INTRODUCTION

STAN Entrance test is a huge thing in Indonesia. STAN (Sekolah Tinggi Akuntansi Negara) or Indonesian State College of Accountancy is favored by many high school graduates as it is one of the credible government-affiliated education institutions. It is a standardized test for test takers who want to go to STAN. In the last ten years, there are at least 40.000 test takers joining the STAN entrance test. In 2016, there are 4490 test takers passing the entrance test of STAN (www.pknstan.ac.id). There is only less than 10 % chance of each test taker to get in to STAN. The entrance test is designed comprehensively in a way to get the best candidate students. The tests is staged in three main tests; Written Test (General Knowledge and English Proficiency Test), Health and Physical Test, and Interview. If the test takers fail in one test, they are not able to continue to the other test. These tests are challenging to some but burdensome to most test takers.

Main focus of this study is the English Test. It is a proficiency test which is going to "measure people's ability in a language, regardless of any training they may have had in that language" (Hughes, 2003, p. 11). In this case, the test would measure the grammar and reading comprehension knowledge of the test takers. The format of cloze items is the test taker must select a correct answer that best suits in a blank in a passage. The test is divided into four main parts: Structure and Written

Expression, Cloze Question, Error Recognition and Reading Comprehension. The test takers are to answer in a separated computerized sheet by blackening one of the options. The English Proficiency Test is aimed at knowing the level English proficiency of the test takers. For the institution, English is included as part of the entrance test as STAN has set a standard of English ability and it is expected that the candidate student of the college have meet the standard of English required. The test takers are highly encouraged to answer as many correct answers as possible as this will affect the national-rank of test takers and will be a consideration for the next phase of test.

Research discussing STAN English Proficiency Test (EPT) is limited and remained unpublished. Confidentiality might be one of the issues why there were not many discussions about the test published. Therefore, this study will evaluate comprehensively STAN EPT using the criteria of the evaluation by Brown & Abeywickrama (2010). Five proposed criteria for 'Testing a Test" used in terms of practicality, reliability, validity, washback and authenticity from Brown & Abeywickrama (2010)'s book. This set of criteria has been used to evaluate some tests for example the test for classroom assessment (Tran, 2012) and in the placement test in Rustaq College of Applied Sciences (Al-Adawi & Al-Balushi, 2016).

Critical Evaluation of the STAN English Test

Practicality

Practicality in a test is related to instruments of logistic and administering in test, including the creating, running and scoring it (Brown & Abeywickrama, 2010). The English Proficiency test in STAN entrance test is most likely to be practical because it met the instruments of Brown and Abeywickrama. STAN EPT administered the multiple choice test. Multiple choice tests are viewed alternatively practical for a large scale testing (Arguelles, Pablo-Lerchundi, Herradon Diez & Banos Exposito, 2011 in Alvarez, 2013; Kumazawa, 2016).). It is not known about the designing of the questions, who involved in designing and how long it took for designing the test; however, the test questions are always revised annually, to provide its validity (Brown & Abeywickrama, 2010). After that, the test committee provided the copies of set of questions with the number of test takers.

In administering the test, it is also practical because it took only 50 minutes for students to do the test. And since the test used computerized answer sheet, its practicality seemed clear in checking and scoring 40.000 answer sheets. Alderson and Bachman (2006) suggested that computer is able in better analyzing and measuring the examinees' response than human raters. Practicality in scoring the test is inevitably important because the result was used to determine whether the test takers could continue to the next test or not. Overall, in the terms of practicality, STAN EPT is practical.

Reliability

Reliability of a test relied most on the consistency and dependability (Brown & Abeywickrama, 2010). Weir (1993) added “for a test to be valid, it must also be reliable”. The reliability of STAN EPT is analyzed in the view of student, rater, test administration and the test itself.

Student-Related Reliability

Students’ reliability issues, according to Brown and Abeywickrama (2010), may relate to the physical and psychological factor of the test takers. In STAN EPT, the pressure would be high as there were thousands of test takers were after the school. Like other standardized test, test takers might have done much preparation and it made them be more fatigue, nervous or overwhelmed. In worst case scenario, they might be sick. This could affect the reliability of the test. In this case, the reliability is low.

Rater Reliability

The STAN EPT used computerized answer sheet. The students’ answers would be checked and rated by the computer. Each set of the test had serial number to respective test takers. It is mandatory to put the serial number in the computerized answer sheet otherwise the answer sheet would not be processed. It is not clear whether the serial number identified the set of questions used, therefore the program to score was also different, but since the computer that would rate and score the test takers work, it tended to be more reliable. Problems of inexperience, inattention or bias in inter-rater and intra-rater mentioned by Brown and Abeywickrama (2010) would be unfeasible as the computer would do the scoring and rating.

Test Administration Reliability

The test administration related to the condition in running the test. From year to year, STAN is well-known for its consistency in the test administration. That is why the number of test takers is grading each year. The test takers’ identity is one important thing to be checked by the room instructor to make sure that it is the test taker himself/herself did the test. The set of the questions and answer sheet is distributed based on the serial number. It is not clear whether the serial number distinguished the questions. The room instructor would read the instructions to test takers, and give opportunities for questions related to the instructions. The instructor would set 50 minutes time to start the test and recall 5 minutes before it is over. Test takers started and finished at the same time. Its strict and consistent rule during the test would minimize any cheatings that could possibly happen.

The consistency in the scoring and marking of the test also became one of the criteria in order to call a test is reliable, which discusses on whether or not the test is consistently marked to the same standard. STAN EPT is scored using computer. As technology has a role in the assessment of language (Chapelle & Douglas, 2006), computer is presented to help faster assess the work of the test takers and to make sure the reliability in the scoring.

Test Reliability

The Test used the multiple choice type of questions. It is considered an objective test as it resulted in fixed responses from the test takers (Brown & Abeywickrama, 2010). Moreover, scoring system is administered in the STAN EPT. Test takers are demanded to have at least one third of the correct answers in order to pass the test. The passing grade system in STAN EPT would make this test challenging. It somehow 'forced' the test takers to over their limit if they wanted to pass the test. In other words, the test reliability of STAN EPT is high.

Validity

Amongst all of the criteria of assessing a test, validity is the most complex and argued to be the most important principle (Gronlund, 1998) in Brown & Abeywickrama, 2010). When the test is designed to assess what is intended to be assessed, the test is valid (Hughes, 2003). The test should not measure irrelevant variables of previous knowledge of a subject, in order not to be called invalid. The validity of STAN EPT is evaluated in terms of content, criterion-related, and face validity.

Content Validity

The representation of language skill learned in the class constructed in a test would indicate validity in content (Hughes, 2003). The test takers are graduates of high school who might come from different type of high schools. In their previous study, they Studied English based on their major. The English test questions were also designed based on their major. The curriculum of teaching and learning for each type of school is different. This is a challenge in designing the test. In STAN EPT 2015, the contents in the reading passages were varied. The topics were from Government, Science, History, and Technology. As STAN test is conducted annually, the content has always been revised, so the content is valid to be tested (Brown & Abeywickrama, 2010).

Viewing from the varied background of test takers and the varied reading passages in STAN EPT, it is less valid in content, even though there is an effort as it tried to accommodate and generalize the English knowledge of the test takers. The most logical reason might be because it is not possible to design the test to only one particular test takers background. The test designed in none specific area; therefore it took some majors' English knowledge. Based on this, the content-validity in STAN EPT is low.

Criterion-related validity

Criterion-related validity deals with the certain criteria reached by test taker from the test (Brown & Abeywickrama, 2010). More on this, the subdivision of this type of validity is the concurrent validity (the actual language proficiency) and predictive validity (test takers readiness or to place the test takers). Since STAN EPT applied the passing grade scoring, in terms of predictive validity, they are 'somehow' placed in a passed or not-passed criterion. However, it did not really prove the concurrent validity, as some test takers might be excellent in Speaking skill, but low in grammar. They would struggle harder for the test. It is different from those who are excellent in grammar but not so good in speaking. They probably had higher chance to pass the test. This condition did not really show the actual language proficiency of the test takers. To sum up, the criterion-related validity in STAN EPT is low.

Face validity

STAN EPT is a direct test. Hughes (2003) mentioned a good direct test would test. It tested the English knowledge in Grammar and Reading comprehension of the test takers. And a direct type of test provides high face validity (Davies, et.al., 1999). A face-validity test is well-constructed, allotted time limit and has clear directions (Brown & Abeywickrama, 2010). As STAN EPT is a standardized test, the construction, instruction and set of time limit have been tested from time to time. In the time constrain, test takers should answer 60 questions in 50 minutes. It means, it would take less than one minute to answer one question. There is an issue of unfairness. However, Brown and Abeywickrama also mentioned that the face validity provides a difficulty level that presents a reasonable challenge. This resulted in a 'biased for best' test. Cohen (2006) mentioned that a 'biased for best' test would result in not only for testing the test takers' ability, but it tended to also go beyond it. It tested the test taking strategies and awareness of the test takers. In a way, the test is designed to encourage the students to perform best, for example, to correctly answer all the questions in 50 minutes. The STAN English Test is viewed as a 'biased for best' test. It intentionally filtered for the best candidates. In the face validity, STAN English Test is high.

Washback

Hughes' defined washback as "the effect of testing on teaching and learning" (2003, p. 1), In a large scale test, like STAN EPT, the washback is evaluated in terms of how prepared the test taker is for the test (Brown & Abeywickrama, 2010). The effect of the testing to the test takers could be viewed from the result of the test. If the result met the criteria of the test, it suggested that the test taker is prepared.

On the other hand, test takers who failed the test would be assumed to have less preparation. It has been discussed in the students-reliability that many factors could affect the test takers performance during the test. For example, if the test taker is sick, she/he would be not in her/his best performance for test and it might affect the result of the test. STAN EPT gave a limited washback to test takers. It did not give feedback in a way for test takers to progress in the future test. Based on this evaluation, STAN EPT is evaluated low in washback criteria.

METHOD

Elder (2016) referred authenticity to represent the connection between a language test and the relevant non-test setting. Brown and Abeywickrama (2010) stated that one condition a test is authentic is if contains language that is as natural as possible. However, they also argued that many test item types fail the real world task. The attempt to assess grammatical or lexical knowledge of the test takers might result in the lack of accommodation towards the authenticity of the test. In STAN EPT, the language is not viewed natural as it is rigid because the test focused on the grammar and how far test takers comprehend a reading text. Therefore, STAN EPT did not fulfill the authenticity criteria.

FINDINGS AND DISCUSSION

STAN EPT is administered for test takers as part of STAN Entrance Test. Using Brown and Abeywickrama (2010)'s criteria, the test has been analyzed critically. STAN EPT is practical and reliable but it has not accommodated the criteria of validity, washback and authenticity. The summary of the analysis will be presented in table 1.

Table 1. Analysis of STAN EPT

Criteria	Sub-criteria	Result
Practicality	-	High
Reliability	Student-Related	Low
	Rater	High
	Test Administration	High
	Test	High
Validity	Content	Low
	Criterion-related	Low
	Face	High
Washback	-	Low
Authenticity	-	Low

CONCLUSION

Based on the study, it is clear now that STAN EPT does not meet the criteria from Brown and Abeywickrama (2010). Therefore, it is important to note that, in test design, it is unlike to have a perfect test. One test might be high in practicality and reliability but low in the other criteria. Take an example from STAN EPT, despite its practicality and reliability is high, its validity, washback and authenticity is still low. This test is imperfect, yet, more and more students are taking the test each year.

In addition, future test design for STAN EPT may consider including spoken test and writing as part of its assessment. Since the test only covers grammar and reading, it is not enough to categorize one is fluent at English. Finally, for the future research, it is recommended to do a comparative study of test evaluation for English test that are used for the students selection by Indonesian government, such as STIS (Sekolah Tinggi Ilmu Statistik), STPDN (Sekolah Tinggi Pemerintahan Dalam Negeri), and other government-affiliated school. This study may help to find out whether the result of test evaluation is similar within the government-affiliated schools.

BIBLIOGRAPHY

Al-Adawi , Sharifa S. A. & Al-Balushi, Aaisha A. K. (2016). Investigating Content and Face Validity of English Language Placement Test Designed by Colleges of Applied Sciences. *English Language Teaching,* 9 (1) 107-121.

Alvarez, Irina Arguelles. (2013). Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice test. *International Journals of English Studies*, 13 (2) 21-38.

Brown, H.D &Abeywickrama, P. (2010). *Language Assessment Principles and Classroom Practices.* NY : Pearson Education.

Cohen, Andrew D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3 (4) 307-331.

Davies, A., Brown, A., Elder, C , Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Hughes, Arthur. (2003). *Testing for language teachers*. Cambridge, New York : Cambridge University Press.

Kumazawa, Takaaki. (2016) Factors affecting multiple-choice cloze test score variance: A perspective from generalizability theory. *International Journal of Language Studies*, 10 (1) 15-30

Weir, C. (1993). *Issues in language testing, Understanding and Developing Language Tests*. New York: Prentice Hall