A Critical Analysis of the STEP Eiken Test's Validity and Reliability.

Stuart Benson

Abstract

This paper will critically analyze Japan's most widely used English testing program, the STEP Eiken test (STEP, 2010). The analysis will discuss relevant background knowledge such as its purpose in Japan and internationally. The test design and procedures involved when participants carry out the test are critically examined. This involves the various formats and levels that are present in the test, the scoring procedures that are given to the participants and whether the information about the test that is given to the 'stakeholders' is adequate.

Finally, the quality of the test in terms of its validity and reliability, specifically its construct, content and face validity are discussed. Unfortunately, very little research on validity has been conducted on the test and even the few studies completed have not been published. Therefore it is difficult to accurately conclude whether the Eiken test has sufficient validity.

Introduction

The STEP Eiken test is an English proficiency test that is administered at 18,000 locations throughout Japan and 45 other countries, three times a year (STEP, 2010). In 2009, approximately 2.3 million people took the test (STEP, 2010). It is regarded as the most widely used English proficiency test in Japan (STEP, 2010). Unlike other proficiency tests such as TOEIC and TOEFL that utilize one single test with a

converted scale-score system, the Eiken test has seven different levels or ‘grades’ of a pre-determined proficiency level which uses a pass/fail system (Nielson, 2000).

The grades range from grade 5 (beginner) to grade 1 (advanced), with two bridging grades (pre-1 and pre-2) (STEP, 2010). As the test has seven different levels with varying requirements, there may be varying measures of validity. This analysis will primarily concentrate on the Pre-1 test which is classified as an ‘advanced’ level (STEP, 2010).

Eiken tests measure the proficiency level of reading, listening, writing and speaking. The speaking section however only occurs from grade 3 through to 1. From grade 3 to grade 1, the test is administered in two stages with reading, writing and listening held at one time then if the participant passes the first stage; they are then required to take the speaking test at a different time.

The first stage of the test is predominantly formed with multiple choice questions. The writing section of the first stage however is a short passage of 100 words (STEP, 2010).

As stated on the STEP official website, the purpose of advanced levels are for “high stakes decisions including admissions to English-medium Universities” (STEP, 2010). Currently, 350 Universities across the United States, Canada, Australasia, and the United Kingdom recognise Eiken results for admission (STEP, 2010). Consequently, the Eiken test is internationally recognised and in turn needs to have a high status of validity and reliability. This will be explored later. The purpose of the lower levels in Eiken is to classify a benchmark of recommended English ability for Junior high school and high school graduates primarily in Japan (STEP, 2010).

Stages of the test

The first stage of the Eiken predominantly consists of multiple choice questions with only one short answer question. With using multiple choice questions or selected response assessments, participants do not need to produce the language and are most appropriate for measuring receptive skills (Brown & Hudson, 1998). In turn, STEP has tried to compensate for the unbalance by conducting a speaking test as a second stage. Unfortunately, compared to the 115 minutes that is spent measuring the participants receptive knowledge, the speaking section consists of only eight minutes. This decreases the validity of the test as it is not an appropriate measure of all the aspects of English. According to the 'Eiken can-do list', after passing the second stage of pre-1 level, participants are able to "ask questions and express opinions about the content lectures and presentations" (STEP, 2010). Unfortunately, the second stage tests are not given out to the public so this cannot be confirmed. However, on the STEP website, the pre-1 test structure does not indicate at all that participants are required to ask questions (STEP, 2010). If this is true, then this would drastically decrease the face validity of the test. The issue of face validity will be dicussed later.

As stated above, the Eiken test uses a pass/fail system which ranges from 60% - 70% according to the level (STEP, 2010). For level pre-1, the pass rate is 70% with 67 items being worth either 1, 2 or 14 points in the first stage (STEP, 2010). Critical information on the marking system of the second stage has not been released to the public prior to administering the test.

This in turn decreases the reliability of the test as crucial information for the various stakeholders is not freely available.

Washback

As the test is available to be taken three times a year, many learners spend a huge amount of time preparing for the test. This in turn produces ‘washback’ on the educational system or society (Bachman & Palmer, 1996, p. 34).

There are various factors that indicate that the test has both positive and negative ‘washback’. STEP states that the tests are “explicitly aimed to enhance positive washback” by “maintaining maximum accessibility, and a strong focus on interacting with and understanding the needs of teachers in order to make the content of the tests as relevant as possible to participants” (STEP, 2010). It is true that the accessibility in Japan with 18,000 testing locations and designing a test that is relevant to the participant has seeminly been achieved. It however seems that test items that are in the test are covered within the nationally approved curriculum guidelines and textbooks (Nielson, 2000). This will dramatically affect the ‘washback’ of the test in two respects. One negative aspect is that if test items are covered in the national textbooks in Japan, “its purpose becomes more one of measuring achievement” (Nielson, 2000, p. 83). This in turn would be detrimental to the content validity of the test as even though it is internationally recognised, the test in fact is not appropriately measuring the true proficiency of the participant. The other negative aspect is how preparation courses are being taught for the test. Messick (1996), states that “the move from learning exercises to test exercises should be seamless. As a consequence, for optimal positive washback, there should be little, if any difference between activities involved in learning the language and activities involved in preparing for the test” (Messick, 1996, p. 242). This however is often not the case with textbooks designed to prepare for the test. One such example is that of the preparation text for Eiken level 1 (Akao, 2010). This textbook uses various activities that are not comparible to that of the test structure. These activities include:

- Translation exercises (Japanese – English)
  (English – Japanese)
- Odd one out exercises (Akao, 2010)

Through analyzing the aspect of 'washback' from the test, there are both positive and negative points that have arisen. This will need to be researched more in detail. One possible outcome that is already present from the negative washback is the rate of pass/fail for 'advanced' levels. In 2009, 72,367 learners applied for level pre-1 but only 10,600 passed (STEP, 2010). It is of course understandable that there are many reasons for participants not passing. If the negative washback however was a cause of failing, this needs to be subsequently reviewed.

Test validity

The concept of test validity has changed over recent years. Prior to the 1990s, validity in tests was generally look at by investigating from one of three separate categories: construct, content and criterion-related validity (Bachman, 1990; Kane, 1992; Messick, 1989, 1996).

Recently however, it is generally agreed that validity is a matter of degree which results from the meaningfulness and appropriateness of the uses and interpretations of test scores (Bachman, 2004, 2005; Cizek, Rosenberg, & Koons, 2008; STEP, 2010; Stoynoff & Chapelle, 2005).

On the Eiken website, it states that the test refers to Bachman and Palmer's (1996) "test usefulness" to help make "transparent, principled decisions regarding the optimal balance of various features for each grade" (Bachman & Palmer, 1996; STEP, 2010). As the Eiken test consists of seven varying proficiency levels and the variety of different ages, this framework is compatible to Eiken. The premise of

Bachman and Palmer's (1996) framework is a co-operating use of six test qualities, reliability, construct validity, authenticity, instructiveness, impact, and practicality (Bachman & Palmer, 1996). The framework allows for the differential management of characteristics according to the needs of each level (Bachman & Palmer, 1996). Therefore, with the higher levels in Eiken, "a higher priority is placed on maximizing validity and reliability, often at the expense of practicality" (Bachman & Palmer, 1996, p. 38).

This is acceptable in Bachman and Palmer's (1996) framework. However information on how validity and reliability is maximized is not evident anywhere on their website or other sources on the test. With the lack of evidence, it is difficult to review their comments. This in turn decreases the face validity of the test.

Through reviewing the evidence that is presented on their website, two key weaknesses in the tests validity and reliability have been identified. The first weakness that is present is the vagueness of what a participant 'can do' if they are able to pass a certain level (STEP, 2010). Each level has numerous 'statements' of what can be interpreted from the results of the test (STEP, 2010). These statements are vague in what it indicates. Specific information is not given and through reading the Eiken 'can-do' list, varying differences in the two advanced levels is difficult to comprehend. This decreases the content and construct validity and in turn the reliability of the test as it is uncertain to what extent the results of test can be interpreted for a particular proficiency level.

The second weakness that decreases the validity of the test is the lack of evidence of validity through validation projects. The website indicates that recently, the approach of 'validity argument' has been discussed by many researchers (Bachman, 2004, 2005; Messick, 1994) and in turn, the Eiken test needs sufficient evidence to justify the interpretations of test scores (STEP, 2010). In turn, various studies

are currently being carried out "to collect evidence to a wide audience, not just for educators and learners in Japan, but including international educators, researchers, and testing specialists" (STEP, 2010). This statement indicates that currently, the Eiken test may not have sufficient evidence of validity for the test to be internationally recognised. This in turn is worrying as stated above, 350 universities around the world recognise the Eiken test as a sufficient proficiency test for admission to Universities. Although there are numerous studies that are presented on the website, few have been publically published.

Conclusion

The lack of sufficient evidence on the validity and reliability of the test has hindered a conclusive result. However, through the various studies that are available, validity of the test is currently lacking. As various studies are currently being undertaken for validating the test, sufficient evidence may be available to conclude whether the test is valid for international recognisation in the near future. Unfortunately, the current evidence leads me to believe that the test is not valid or reliable. The most concerning result that was presented in Nielsons' (2000) article was that a staggering 79% of the test items in level three of the Eiken test were not reliable for the test due to discrimination between test takers or are too easy or difficult for the specific level (Nielson, 2000, p. 90). This indicates that the test needs urgent attention for validation.

References

Akao, F. (2010). *Eiken Jun 1 Kkyuu Shuchu Zemi* (2nd ed). Japan: Obunsha.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analysis for language assessment.* Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F., & Palmer, A. S. (1996). Test usefulness: Qualities of language tests. In *Language testing in practice* (pp. 17-42). Oxford: Oxford University Press.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly 32*(4), 653-675.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and psychological measurement, 68*(3), 397-412.

Kane, M. T. (1992). An arguement-based approach to validity. *Psychological bulletin, 112,* 527-535.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language testing, 13*(3), 241-256.

Nielson, B. (2000). Determining Test Reliability and Quality of Eiken Test Items : A Statistical Analysis of First Year Kosen Student Responses to Test Items of an Eiken Third Level Test *Research reports, Kushiro Technical College 34,* 81-91.

STEP, E. (2010). STEP (Soceity For Testing English Proficiency). Retrieved 15/11/, 2010, from http://stepeiken.org/

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing: a resource for teachers and administrators.* Alexandria, VA: TESOL Publications.