ISSN: 2527-8037

Proceedings of the 1st English Education International Conference (EEIC) in conjunction with the 2nd Reciprocal Graduate Research Symposium (RGRS) of the Consortium of Asia-Pacific Education Universities (CAPEU) between Sultan Idris Education University and Syiah Kuala University

November 12-13, 2016, Banda Aceh, Indonesia

DIY: DESIGNING A READING TEST AS RELIABLE AS A PAPER-BASED TOEFL DESIGNED BY ETS

Faisal Mustafa[*1] and **Heri Apriadi**[2]

[1]Syiah Kuala University, Banda Aceh, INDONESIA
[2]Language Center of Syiah Kuala University, Banda Aceh, INDONESIA
[*]*Corresponding author: faisal.mustafa@unsyiah.ac.id*

Abstract

Designing a reliable reading test has been proven very difficult and following test design steps is time-consuming; however, it is very significant both for teachers and test centers. Therefore, this research was aimed at proposing an alternative, time-saving method of designing a reading test to measure EFL student proficiency. The method proposed was to use a standardized test as a template both for texts and question types. To prove that this method was effective, a reading test was designed by using the method and tested for its reliability level, and compared to the reliability of reading test in PBT TOEFL. The results suggested that the test was highly reliable (85%), very close to the reliability level of the PBT TOEFL designed by ETS (86%). In addition, the scores obtained were significantly correlated with the scores achieved in an ETS-designed test (91%). Therefore, it is recommended that test makers follow the procedure of designing a reading test by using standardized test such as PBT TOEFL as the template in developing a reliable test, thus reliability test and revision process can both be bypassed.

Keywords: *Reading test, test design, language assessment, standardized test, TOEFL.*

INTRODUCTION

Information on English proficiency level is continuously required by universities, scholarship providers, and other stakeholders in recruiting (Nurhayati & Gir, 2014, p. 153). English proficiency level is commonly obtained through an English language test. One of the most widely-known standardized English language tests is TOEFL (Test of English as a Foreign Language) (Chen, 2016, p. 65). However, the test is cost-prohibitive, and thus TOEFL-like-test was conducted by government learning institutions, as well as private institutions, such as universities. In Indonesia, TOEFL-like-tests are very popular since many universities require the score both for registration (Nurhayati & Gir, 2014, p. 135; Sugeng, Saleh & Suharto, 2012, pp. 191–192) and graduation (Marzuki, 2008, p. 95; Sabarun, 2012; Salam, Fergina & Suparjan, 2012, p. 15).

TOEFL test materials are designed, and can only be used with paid permission from ETS. TOEFL-like test materials used by other institutions are test samples from TOEFL books (Mufidah, 2014, p. 80; Palandi & Pudyastuti, 2011). Using these materials for tests has some significant drawbacks. First, the level of these tests have been found to be inconsistent with that of tests provided by the ETS (Hilke & Wadden, 1997). As such, scores obtained by the students in this test may not accurately represent the students' level of TOEFL-based English proficiency. Second, since the books are sold to the public, there is a chance that the students taking the TOEFL-like test have encountered the book from which the test was taken. Finally, the books do not provide information regarding the procedure of test design or level of reliability; thus the score obtained by the test takers is unreliable. Therefore, some institutions decided to make their own TOEFL-like test. However, making a test involves time-consuming processes (Bachman & Palmer, 1996, p. 165), especially in piloting and revising (Catelly, 2014, p. 395). Consequently, alternative and faster test design procedure is essential for TOEFL-like test centers.

This research was aimed at proposing an alternative procedure of designing TOEFL test material. Designing *Structure and Written Expression* test section has been proposed by Mustafa (2015); therefore, this paper focused on designing a reliable, standardized level, reading test material.

METHODS

This section describes procedures for test design, delivery and analysis.

Test Developer Requirements

Before beginning test design, it is essential to define eligible test developer for this purpose. There are very few discussions on what it takes to be a language test developer who writes a language test. However, it can be assumed that a test developer should have detail knowledge on language testing (Bachman & Palmer, 1996, p. 158). In addition, a language test should know the language for which the test is made. In the case that the test developer is a non-native speaker of the language, it is safe to suggest that the highest level of language proficiency is required. In Common European Frame of Reference (CEFR), the highest level is termed as proficient user (C levels) (Council of Europe, 2001, p. 24), which can be converted into the minimum TOEFL ITP score of 627, 64 for listening and structure sections, and 63 for reading section (Tannenbaum & Baron, 2011). Since the test materials designed for this research were reading tests, the writers who developed the tests have thus reached the proficient user level in reading comprehension based on their TOEFL scores.

Procedures of Designing a Reliable Reading Test

Designing a test involves several steps proposed by many experts in language testing, as presented in Table 1 (Brown, 2004). The alternative method of designing a test we propose in this paper is based on a work by Mustafa (2015). This method purposes to use existing standardized test materials as a template in order to bypass some time-consuming steps, as presented *Proposed Steps*, Table 1.

Table 1. Procedures of designing a reliable reading test.

No.	*Steps suggested by Brown* (Brown, 2004)	*Proposed steps*
1.	Deciding ”the purpose and objective of the test”	*X*
2.	Deciding what needs to be tested in reading and how the test will look like	*X*
3	Selecting and editing texts	√
4	Writing test items	√
5	Writing test options (answers and distractors)	√*
6	Piloting	*X*
7	Reviewing and revising	*X*

*prioritized step

In designing a reading test, the purpose of the test has been previously determined, thus the first step can be skipped. What needs to be tested and how the test will look will also have been provided by the template, enabling a test developer to eliminate this time-consuming process, which Brown (2004) claimed as the most difficult process in test development.

The first step in developing a reading test from an existing standardized reading test is selecting the template test. In this case, a reading section from the TOEFL was chosen the template. Hilke and Wadden (1997, p. 38) suggest that the reading test should match the template in terms of question types and their frequency, and topic areas of the texts. In addition, Hilke and Wadden (1997, p. 43) added that text length and difficulty should also be considered.

Text Selection

As suggested by Hilke and Wadden (1997), the subject areas, text length, and difficulty level of all texts should match those in the TOEFL. Matching the topic areas and the lengths should be an easy process. The text selected should be understood by general readers without specific knowledge on the topic, and it should not be culturally-offensive or sensitive, such as text involving natural disasters, death, or other tragedies (ETS, 2009, pp. 12–15). It is advisable to choose texts from magazines such as National Geographic since it is intended for general readers. The text length recommended for reading section in the TOEFL is between 300 and 400 words (Cohen & Upton, 2007, p. 213). In case the text is too lengthy, some phrases or sentences which do not contribute to making the text more comprehensible may be deleted. Difficulty level is determined as a function of vocabulary and syntactic coverage (Akbari, 2014, p. 125; Heilman, Collins-Thompson, Callan, & Eskenazi, 2007, p. 465; Miltsakaki & Troutt, 2008, p. 96; Zhang, 2012, p. 572). To analyze those coverages, ETS has designed a tool called TextEvaluator formerly known as SourceRater, which is accessible to the public at https://texteval-pilot.ets.org/TextEvaluator/ (Sheehan, *et al.,* 2010, p. 35). This tool uses corpora, which has been claimed accurate (Cobb & Horst, 2004, p. 15), for vocabulary complexity analysis. The syntactic coverage was analyzed based on ”the average number of clauses per sentence, the average number of words per sentence, and the average frequency of prepositions” (Sheehan, *et al.,* 2010, p. 18). The texts selected for test design in this research were adjusted to the difficulty level of the template texts. When the vocabulary complexity was higher or lower than the templates, some vocabulary was modified and reanalyzed.

Question Types and Frequency

The purpose of a reading test is to assess test takers' skills to answer comprehension questions, such as questions about main ideas (Leopold & Leutner, 2012, p. 20), positive and negative factual questions, implied detail questions, and questions on vocabulary (Lee, 2002, p. 153). Consistent with this, the reading section in the TOEFL test contains questions in the area of main ideas, stated and implied details, and vocabulary meaning (Phillips, 2001, p. 359). An analysis of question type frequency in a randomly selected PBT TOEFL test reveals the frequency of question types as presented in Table 2.

Table 2. Frequency of reading question types in a PBT TOEFL test material.

No.	*Question types*	*Frequency*
1	Main idea	8%
2	Vocabulary	20%
3	Stated detail	22%
4	Unstated detail	10%
5	Reference	14%
6	Inference	22%
7	Locating information	4%

As each question type exhibits different difficulty levels and question type distribution is not uniform, the question types of each text should be adjusted to maintain consistency with the template. According to Cohen and Upton (2007, p. 237), using background knowledge is a common test-taking strategy. As such, questions should also be designed so as not to be answerable based on test taker's background knowledge.

Options (Correct Answer and Distractors)

Writing options for a multiple-choice test is very challenging and involves careful analyses of text contents. The right answers should be a paraphrased version of the words or phrases in the text. Test developer should avoid ”unintentional clues” such as morphosyntax or phonology (Brown, 1996, p. 55). Writing the incorrect options called distractors, is very difficult and requires test developers' careful attention (Nation & Beglar, 2007, p. 12). Distractors are meant to pull test takers away from selecting the right answer when they do not know the answer (Brown, 1996, pp. 54–55). The following is the summary of distractor requirements:

- They should not be apparent to the test takers because the test taker will be able to correctly guess the answer by eliminating the distractors (Cohen, 2012, p. 264).

- No two distractors are a paraphrase of each other because they will be known wrong to test takers (Allan, 1992, pp. 102–103).

To achieve those requirements, the distractor should be verbatim or paraphrased information in the text which can be found close to where the correct answer is located. The distractors might contain untrue information and information which is not actually given in the text.

Proof of Reliability

After designing a reading test, for the purpose of this research only, a reliability test was conducted to prove that the test designed by following the procedures presented above is reliable and matches the difficulty level of the template. The test was piloted for two purposes, reliability test and score comparison analysis. For test reliability, the writers employed test-retest reliability analysis by using the Pearson's Product-Moment Coefficient of Correlation (r) suggested by Best and Kahn (2006, p. 384). The interval between tests was two weeks, with the consideration that the participants, who were selected by using convenient sampling technique, had forgotten their previous answers and, in case they studied, their proficiency would not have upgraded to the level that it would change their scores. 47 participants sat in these tests. Meanwhile, for score comparison analysis, the participants were retested with the template and their scores were compared to what they obtained in the designed tests. 20 participants sat in this third test. This score comparison analysis is required to show that the difficulty level of the designed test matches that of the template.

RESULTS AND DISCUSSION

For the purpose of reliability analysis, the results of two tests in percentage are presented in the following Table 3. *X* in Table 3 represents the total score of all participants in the first test and *Y* the total score in the second test.

Table 3. Test reliability calculation table

$\sum X$	$\sum Y$	$\sum (X-\bar{X})$	$\sum (Y-\bar{Y})$	$\sum (X-\bar{X})^2$	$\sum (Y-\bar{Y})^2$	$\sum (X-\bar{X})(Y-\bar{Y})$
1175	1192	0	0	300	285	239

To calculate the reliability level, the data in the table were inserted into the Pearson's Product-Moment Coefficient of Correlation formula, as presented in the following.

$$r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}} = \frac{239}{\sqrt{300x285}} = .85$$

The calculation resulted in a reliability level of 85%, which suggests that the test design by using a template is highly reliable and thus eligible for use as a standardized test (Frisbie, 1988, p. 29). Compared to the reliability level of the template, i.e. 86% (Wainer & Lukhele, 1997, p. 10), the designed test reliability level is extremely close. However, will a test taker who sits in real TOEFL ITP and this "TOEFL Imitator" obtain similar score? To answer this question, the correlation (r) between the scores obtained when using the designed test (X) and those when using the template (Y) was calculated using Pearson's Product-Moment Coefficient of Correlation formula.

Table 4. Score correlation table between the designed test and the template

$\sum X$	$\sum Y$	$\sum (X-\bar{X})$	$\sum (Y-\bar{Y})$	$\sum (X-\bar{X})^2$	$\sum (Y-\bar{Y})^2$	$\sum (X-\bar{X})(Y-\bar{Y})$
286	239	0	0	240	648	360

The correlation was calculated by using Pearson's Product-Moment Coefficient of Correlation formula.

$$r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}} = \frac{360}{\sqrt{286x648}} = .91$$

The result above confirmed that the scores obtained were very significantly correlated (91%). Therefore, either taking the test designed by using a standardized test as a template or using the template itself, a test taker will get the same score.

CONCLUSION

This paper proposed an alternative, time-efficient procedure of designing a standardized reading test by using an existing standardized test as the template. The procedure proposed includes selecting texts which have similar difficulty with the template, writing questions and answer-options. Other time-consuming processes such as determining construct, piloting, reviewing and revising the test can be bypassed. A test designed by using this alternative procedure was piloted and yielded the reliability level of 85%, which is very high and meets the requirement of a standardized test. In addition, the scores the test takers obtained were highly correlated with their scores when tested with the template test, i.e. 91%. Therefore, to save time and human-resources, it is recommended that a reading test be designed by using a template and following the alternative procedure proposed in this paper to obtain a reliable test.

REFERENCES

Akbari, Z. (2014). The role of grammar in second language reading comprehension: Iranian ESP context. *Procedia - Social and Behavioral Sciences*, *98*, 122–126.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, *9*, 101–119.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Best J. W., & Kahn J. V. (2006). *Research in education* (10th Edition). Boston: Pearson Education, Inc.

Brown, D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall Regents.

Catelly, Y.-M. (2014). Optimizing language assessment – Focus on test specification and piloting. *Procedia - Social and Behavioral Sciences*, *128*, 393–398.

Chen, A. C. (2016). A critical evaluation of text difficulty development in ELT textbook series : A corpus-based approach using variability neighbor clustering. *System*, *58*, 64–81.

Cobb, T., & Horst, M. (2004). Is there room for an academic wordlist in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 15–38). Philadelphia: John Benjamins Publishing.

Cohen, A. D. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.), *The Rouledge handbook of language testing* (pp. 262–277). Oxon: Routledge.

Cohen, A. D., & Upton, T. A. (2007). `I want to go back to the text': Response strategies on the reading subtest of the new TOEFL(R). *Language Testing*, *24*(2), 209–250.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

ETS. (2009). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.* Princeton, N. J.: ETS. Retrieved from https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping the TOEFL® ITP Tests onto the Common European Framework of Reference*. Princeton, N. J.: ETS. Retrieved from: https://www.ets.org/s/toefl_itp/pdf/mapping_toefl_itp_scores_onto_the_common_europea_framework_of_reference.pdf

Frisbie, D. a. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25–35.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Human Language Technology Conference* (pp. 460–467). Rochester, New York.

Hilke, R., & Wadden, P. (1997). The TOEFL and its imitators: Analyzing the TOEFL and evaluating TOEFL-prep texts. *RELC Journal*, *28*(1), 28–53.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement*, *39*(2), 149–164.

Leopold, C., & Leutner, D. (2012). Science text comprehension: Drawing, main idea selection , and summarizing as learning strategies. *Learning and Instruction*, *22*(1), 16–26.

Marzuki, D. (2008). Keterampilan Reading TOEFL like test mahasiswa semester V Jurusan Akuntansi Politeknik Negeri Padang. *Jurnal Akuntansi & Manajemen*, *3*(2), 95–106.

Miltsakaki, E., & Troutt, A. (2008). Real-time web text classification and analysis of reading difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 89–97). Columbus, Ohio: Association for Computational Linguistics.

Mufidah, N. (2014). The English teachers' mastery in TOEFL Prediction. *Journal on English as a Foreign Language*, *4*(2), 79–86.

Mustafa, F. (2015). Using corpora to design a reliable test instrument for English proficiency assessment. In *Teaching and Assessing L2 Learners in the 21st Century* (pp. 344–352). Denpasar.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nurhayati, I. K., & Gir, R. R. W. (2014). Analisis perbandingan nilai TOEFL dengan nilai mata kuliah Bahasa Inggris mahasiswa [Analysis of comparison between TOEFL score and the score for English module]. *Jurnal Sosioteknologi*, *13*(2), 134–146.

Palandi, J. F., & Pudyastuti, Z. E. (2011). Desain sistem TOEFL Prediction untuk membantu persiapan tes TOEFL [Designing TOEFL PRediction software to hep with TOEFL test preparation]. *SNATIKA*, *1*(1).

Phillips, D. (2001). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests*. New York: Pearson Education.

Sabarun. (2012). The students' scores on the different Institutional TOEFLs at the sixth semester English Department students of the Palangka Raya State Islamic College. *Educate*, *1*(2).

Salam, U., Fergina, A., & Suparjan. (2012). Kebijakan TOEFL di Universitas Tanjungpura: Analisis studi kasus [TOEFL policy in Tanjungpura University: A case analysis]. *Jurnal Guru Membangun*, *28*(2), 13–25.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards*. New Jersey.

Sugeng, B., Saleh, S. M., & Suharto, G. (2012). Penguasaan Bahasa Inggris mahasiswa baru UNY Tahun Akademil 2005/2006 - 2009/2010 pada kriteria TOEFL-Like. *LITERA*, *11*(2), 189–203.

Wainer, H., & Lukhele, R. (1997). How reliable is the TOEFL test? *ETS Research Report Series*, *1997*(1), i-23.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*(4), 558–575.