Distinguishing TOEFL Score: What is the Lowest Score Considered a TOEFL Score?

Faisal Mustafa[1]* and Samsul Anwar[2]

[1]*Department of English Education, Faculty of Teacher Training and Education, Syiah Kuala University, Banda Aceh, Indonesia*
[2]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Syiah Kuala University, Banda Aceh, Indonesia*

ABSTRACT

Paper-based TOEFL scores have been used to determine the level of English proficiency for EFL learners for various purposes. However, in repeat tests some lower scores fluctuate despite no additional classroom learning, thus they cannot be used to judge the English level of those taking the test. There is limited research into the lowest score that does not fluctuate outside the Standard Error of Measurement, which the Educational Testing Service (ETS) set at 13 points. Therefore, this research was aimed at determining the lowest score which can be used for distinguishing the students' learning progress or proficiency. Scores of 1,180 test takers who took paper-based TOEFL a minimum of three times over three days to two weeks were analyzed statistically. The analysis revealed that the scores stopped fluctuating outside the Standard Error of Measurement when test takers reached the score of 417. Therefore, not until a test taker obtains the minimum paper-based TOEFL score of 417 can their English level be determined by the TOEFL score. This research has significant implications for employers, universities and high schools that currently use a TOEFL score lower than 417 as the minimum entrance or graduation requirement.

Keywords: TOEFL, lowest real TOEFL score, minimum score, placement test, Standard Error of Measurement

INTRODUCTION

English proficiency tests are designed to measure the level of English for various purposes. Paper-based TOEFL is one such test and is the most preferred English proficiency test because it is accepted by many institutions (Brown, 2004, p. 84). One reason for its popularity is that it is easy to obtain and create (Mustafa, 2015;

Mustafa & Apriadi, 2016). Test takers only need two hours to complete the test. In addition, TOEFL scores are also used as placement tests to indicate the progress of learning (Brown, 1996, p. 12), for job and scholarship application requirements, and university enrolments (Bachman & Palmer, 1996, p. 185). However, according to probability theory, the probability for each question being answered correctly by random guesses, considering the questions are multiple choice with four options, is 25%, or TOEFL scores between 323 and 363 (Allan 1992). In addition, experience indicates that scores greater than 363 fluctuated when a test taker took multiple tests without any preparation in between. However, ETS, the TOEFL test developer, does not warn the score users about this weakness. Moreover, although there has been much research into paper-based TOEFL, none addressed the issue of score fluctuation. As a result, researchers such as Sabarun (2012), had used TOEFL to categorize students with the scores of 350 and 370 into two different levels. In addition, Heffernan (2006, p. 165) considered the changes in TOEFL scores obtained by undergraduate university students in Japan between pretest and post-test of 340-393, 347-400, 363-390, and 387-397 as improvements. Therefore, it is essential to figure out what is the lowest score which can be used in determining students' learning progress or placing students into different group levels. The current study aimed at finding out this score by utilizing statistical analysis. The result is significant for institutions which use PBT TOEFL score as criteria in recruitment, placement, and admission, or other requirements.

LITERATURE REVIEW

This section discusses variables involved in this research, i.e. TOEFL and fluctuation in TOEFL scores, reliability and Standard Error of Measurement for TOEFL.

Test of English as a Foreign Language (TOEFL)

TOEFL is one of the standardized language tests for foreign language learners. It is a reliable test designed by Educational Testing Service (ETS) based in New Jersey, U.S.A. The test has evolved from a paper-based test to an internet-based test through several phases of revision. It was first used as a paper-based test in the early 1960's (Spolsky, 1990). The test is in three sections, i.e. listening comprehension, structure and written expression, and reading comprehension. In 1998, a computer-based TOEFL (CBT TOEFL) was developed, which included Test of Written English (TWE), but is now discontinued, replaced by the internet-based TOEFL (iBT TOEFL) (ETS, 2011a, pp. 3-5). The iBT TOEFL offers both English written and spoken tests, while structure and written expression, which were tested in PBT and CBT, has been excluded in the iBT TOEFL (ETS, 2005, p. 4). Although an internet-based test is very effective, it is not possible in the areas where internet connection is unavailable or unreliable, and therefore PBT TOEFL with TWE is an alternative (ETS, 2011, p. 3).

In addition, iBT TOEFL is not required in many universities in non-English speaking countries due to its unaffordability. In that case, Institutional TOEFL, which is PBT TOEFL without TWE (Tannenbaum & Baron, 2012, pp. 7-8), is the alternative.

TOEFL as A Type of Language Assessment

Teachers have been using language assessments to judge the success of both teaching and learning practices (Brown, 2004, p. 4). A test, as a subset of assessment, is used to measure language proficiency (Alderson, 2007, pp. 22-25), as required for placement in a language training, scholarship or job application. One such test is paper-based TOEFL (Brown, 1996, p. 5), a type of English proficiency test which is very popular among EFL learners. It was first introduced in the United States in 1963 (Wainer & Lukhele, 1997, p. ii). Although it is claimed that the test is a valid measure of nonnative speaker English proficiency (Rosenfeld, Oltman, & Sheppard, 2004, p. 1), some have argued that the test does not represent the whole language performance (Chalhoub-Deville & Turner, 2000, p. 537). One criticism was that communicative performance was not tested in paper-based TOEFL. In addition, Institutional Testing Program TOEFL (ITP TOEFL), "a retired version" of paper-based TOEFL (Nisbet, 2002, p. 31) administered for educational institutions to make admission decisions or as a graduation requirement (Takagi, 2011, p. 113), does not test either communicative or written English performance. ETS responded positively to the feedback from these researchers and the new theories in language testing and thus revised the TOEFL to include all components of language performance (ETS, 2010). The upgraded version is known as internet-based TOEFL. Test takers admitted that it is a more representative tool to measure proficiency in English for Academic Purposes (DeLuca, Cheng, Fox, Doe, & Li, 2013, p. 673). However, paper-based and ITP TOEFL are still used today when iBT TOEFL is not possible, for example, as a result of unavailable internet connection and cost restrictions.

Fluctuation in TOEFL Scores

Like other multiple-choice tests, the reliability of paper-based TOEFL is threatened by random guesses. There are four options for each question which means that the possibility of guessing correctly is 25%. Table 1 presents the scores resulting from purely random guesses.

Table 1
TOEFL scores resulting from random guesses

No	Section	Correctly guessed	Scores
1	Listening	13	36-40
2	Structure	10	30-35
3	Reading	13	31-34
	Total	36	323-363

Note: The conversion is based on Gear and Gear (1996)

Table 1 above shows that a test taker relying on guesses can obtain a score between 323 and 363. However, Tannenbaum and Baron (2012, p. 14) categorized these

scores as level A2 in the Common European Framework of Reference (CEFR), which according to Council of Europe (2001) the students have the ability to

- understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment);

- communicate simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters;

- describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need. (p. 24).

Another threat to test validity and reliability is what Thorndike (1951, p. 568) referred to as "test-wiseness" strategy, the ability to answer a multiple-choice test correctly without having the knowledge required to answer the question (Millman & Bishop, 1965, p. 707). According to Allan (1992), the strategies include: a) absurd option, b) grammatical cue, c) item give-away, d) longer length option, e) option inclusion, f) precise option, g) similar option, h) choose neither or both of two options which imply the correctness of each other, i) choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other, j) specific determiner, and k) stem-option. (p. 101)

Research by Tavakoli and Samian (2014) revealed that test takers used test-wiseness strategy in paper-based TOEFL. Yang (2000) analyzed Listening and Reading Comprehension Sections in one of the TOEFL materials and discovered that 48% to 64% of questions across the sections were "identified as susceptible to test-wiseness." Allan (1992, p. 108) provided the average number of correct answers which can be obtained by using test-wiseness strategy, i.e. 55%. Table 2 shows the scores which can be obtained by using test-wiseness strategy.

Table 2 shows that the minimum paper-based score obtained by using test-wiseness strategy is 323 and the maximum is 407. These scores consider the percentage of questions susceptible to test-wiseness, which ranges from 48% to 64% of the questions. Since there is 55% chance of correctly answering the susceptible-to-test-wiseness questions, the number of such questions was multiplied by 55%.

These two threats to validity and reliability result in fluctuation in TOEFL scores when the test is repeated. Random guesses and test-wiseness strategy are used less often by high proficiency groups (Ebel, 1968, p. 321; Kashkouli & Barati, 2013, p. 1584). This suggests that low proficiency test takers need to rely on their test-taking strategy, and those with zero-knowledge should take random guesses. In addition, the scoring system for TOEFL does not give a penalty for incorrect answers, which produces a bias for low proficiency students due to guessing (Reid, 1977, p. 335). Guessing can be right, or wrong, as can the

answers based on test-wiseness, producing fluctuation when a test taker repeats the test.

Table 2
TOEFL scores resulted from test-wiseness strategy

No	Sections	Susceptibility of 48%		Susceptibility of 64%		Scores	
		Questions	Correct (55%)	Questions	Correct (55%)	48%	64%
1	Listening	24	13	32	18	35-40	43-44
2	Structure	19	11	26	14	30-35	36-38
3	Reading	24	13	32	18	31-34	38-40
	Total					323-363*	390-407*

Note: The conversion is based on Gear and Gear (1996)

Reliability and Standard Error of Measurement for TOEFL

Reliability refers to the consistency and accuracy of measurement when a test is "administered under similar conditions" (Hatch & Lazaraton, 1991, p. 530). The reliability level ranges between 0% and 100%. When a test is re-administered to a group of participants more than once, and they obtain exactly the same scores, the reliability of the test is 100%. For classroom use, Douglas (2010, p. 107) and Wells and Wollack (2003, p. 5) suggest a reliability level of 70%. For standardized tests such as TOEFL or IELTS, the level should not be less than 85% (Frisbie, 1988, p. 29). Among other types of reliability test, Hatch and Lazaraton (1991, p. 531) ranked test-retest method as the most preferred. It is calculated by looking at the correlation between the first and the second test (Douglas. 2010, p. 105), with the following formula from Best and Kahn (2006, p. 384):

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Where y = sum of second test subtracted from each second test score

xy = sum of each x multiplied by each y

For paper-based TOEFL, the reported reliability for an overall score was 96%, 93% for listening comprehension section, 90% for structure and written expression section, and 88% for reading comprehension section ETS (2016). Therefore, paper-based TOEFL is considered a reliable test.

The reliability level allows us to determine the range of fluctuation if the test is repeated, known as Standard Error of Measurement. It is calculated by using the following formula proposed by Douglas (2010, p. 108):

$$SEM = SD\sqrt{1 - Rel}$$

Where

SEM = Standard Error of Measurement.
SD   = Standard Deviation
Rel. = Reliability

ETS (2016) reported that the Standard Error of Measurement for Paper-based TOEFL is 13 points. Therefore, if the score obtained by a test taker reflects his English proficiency, the fluctuation of his score will not be larger than 13 points when he repeats the same test.

METHODS

This section presents description of the sampling procedure, data collection and statistical analysis.

Study Design, Population and Sampling Procedure

This study used TOEFL scores collected from the Language Center of Syiah Kuala University, the oldest and largest university in Aceh, the westernmost province of Indonesia. The test was administered by the Center as a graduation requirement for students, who were required to obtain a minimum score of 450, as well as some members of the public who took the test for job and scholarship applications. Others took the test for their self-assessment and practice. The test material used was a reliable standardized TOEFL design by ETS. The raw scores are converted to scaled scores using a statistical method called Item Response Theory (IRT) with a 3PL Model (Way & Reese, 1991, p. 18). This method requires values for item discrimination and item difficulty, which are not revealed by ETS to public. Therefore, it is less likely to use the formula to convert the scores. Thus, a conversion table should be used. Conversion tables that are easy to use are provided by Phillips (2003) and Pyle and Page (1995). The table provided by Phillips (2003) is preferred due to the popularity of the book in which the table is provided. Moreover, the conversion tables are very similar. The data were collected between 2011 and 2016. In order to examine the fluctuations in test scores, this study used the data from test takers who sat the test at least three times, and for the test takers who took the test three times, the middle test was used as the baseline of the dataset. The absolute difference between the first test and the baseline was calculated, as well as the third and the baseline. These absolute differences were used to measure the fluctuation of the TOEFL scores. The absolute difference between the first and the baseline test was measured as the lower deviation, while the absolute difference between the third and the baseline test was measured as the upper deviation. According to the study design, 45,000 TOEFL scores were taken from 10,850 test takers who took the test at least three times. For the test takers who took the test more than three times, the first three tests were used as a dataset with the second test as the baseline. Furthermore, the second to the fourth tests were also used as a dataset, but the baseline was shifted to the third test as the middle test between the second and the fourth test. The same procedure was used for other numbers of times the

test was taken. Therefore, test takers who took the test three times contributed only 1 dataset, those who took the test 4 times gave 2 datasets, 3 datasets for the test takers taking the test 5 times, and so forth (Figure 1). The time between baseline and the other tests was restricted to three days to two weeks as a sampling criterion. This time lag was decided based on the research result by Kokhan (2012, p. 303) who suggests that the TOEFL scores tend to be less stable as the interval gets longer. The minimum interval of three days was used because no test takers repeated the test in less than three days.
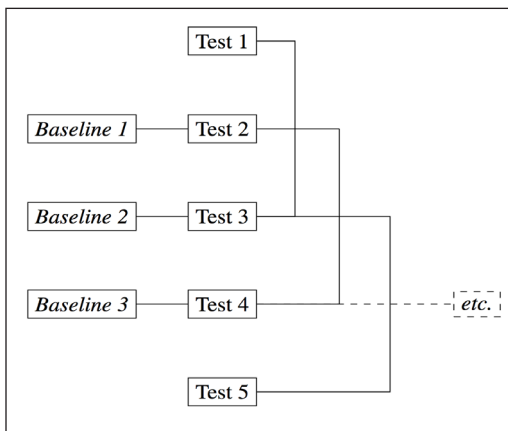
Statistical Analyses

The study aimed to determine the lowest TOEFL score where the fluctuation is no larger than 13 points, the Standard Error of Measurement of paper-based TOEFL given by ETS (2016). In order to achieve this objective, the baseline score was used as the sample score in the study. These datasets were examined by One-Sample T-Test for lower and upper deviations to test whether the mean of the deviations is equal to or less than 13. Figure 2 in the following provides a clear description about the deviations.
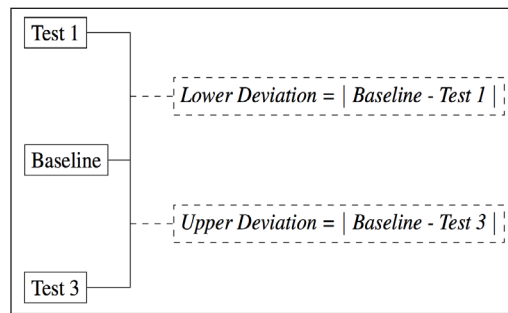

Figure 1. Baseline in a dataset for test takers taking more than three tests

Figure 2. Calculating upper and lower deviations in a dataset

In Figure 1, the first dataset consists of *Test 1*, *Test 2* and *Test 3*, where *Test 2* is the baseline, considered as a sample score to be evaluated. The second dataset includes *Test* 2, *Test 3,* and *Test* 4, and now *Test 3* is the baseline, and so forth.

One of the most important assumptions of the Independent T-Test is that the data should be normally distributed. A Shapiro-Wilk test was used to test the normality assumption at 2.5% of significant error. The baseline scores for evaluation were set from the lowest to the highest possible scores, ranging from 310 to 677. The score of 310 was used as the lowest score because ETS (2011, p. 14) claimed that 310 is the lowest observed score obtained by test participants. After all possible scores had been examined,

the results were compared for all scores starting from 310. The first baseline score for which both lower and upper deviations of the T- Test result were not significant (P>0.05) and the Shapiro-Wilk test were not significant (P>0.025) was the score where the fluctuations were lower or equal to 13 points. This score was considered the lowest score which can be used to distinguish the level of English proficiency.

RESULT

Statistical description revealed the characteristics of the scores in the population, i.e. minimum, maximum, median and mean scores. For the first test, the extreme values - minimum and maximum scores - were 217 and 627 respectively. The median and mean scores of the first test were 363 and 369.3. The median indicated that the scores

obtained by 50% of test takers in the first test were below 363 and those received by the rest of the test takers were greater than 363. The average was 369.3 with 95% confidence interval, i.e. 368.68 and 369.95. The second test (baseline) had the extreme values of 217 (minimum) and 617 (maximum). In the second test 50% of the samples obtained scores below 367, and the rest were above that score. The average score was 373.03 with 95% confidence interval, i.e. between 372.37 (lower bound) and 373.69 (upper bound). Finally, the extreme values for the third test were scores of 217 and 620, with the median score of 370. The mean score, with 95% confidence interval for the third test was 376.94, with 376.24 for the lower bound and 377.64 for the upper bound. The summary is presented in Table 3.

Table 3
Descriptive statistics

Test	n	Min	Med	Max	Mean	95% Cl of Mean	
						Lower bound	Upper bound
1st test	15,000	217	363	627	369.32	368.68	369.95
2nd test	15,000	217	367	617	373.03	372.37	373.69
3rd test	15,000	217	370	620	376.94	376.24	377.64

Table 3 shows that on average the mean, lower and upper bounds would be likely to increase by around 3 points every time the test takers retook the test. This increase is presented in Figure 3.

The scores in Figure 3 above, however, do not have any meaningful interpretation in this study because the last possible numbers in TOEFL scores based on the TOEFL

scoring system are 0, 3, and 7 (e.g. 360, 363, 367, 370, ...). Therefore, the mean scores must be rounded to 370 (first test), 373 (second test/baseline) and 377 (third test).

In order to find out which score had lower and upper deviations within 13 points, we performed a One- Sample T-Test for each score, starting from 310. Our null hypothesis states that a score with average deviations
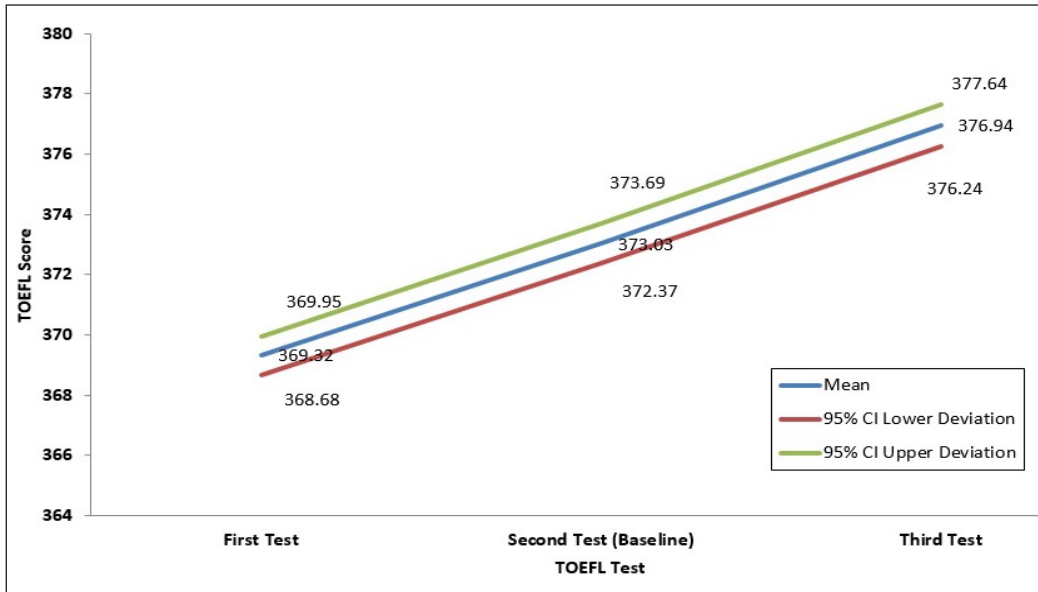
Figure 3. Mean scores and their 95% confident intervals

lower than or equal to 13 points for takers who took two consecutive tests within two weeks is considered a real TOEFL score which represents the test taker's English proficiency. Among the whole population, 1,180 test takers (7.87% of the population) met the sampling criteria. We examined both their lower and upper deviation, ranging from 310 to 677. However, the examination was stopped at 457 due to absence of the required number of samples for conducting a One-Sample T-Test. The number of samples for each baseline is shown in Figure 4.
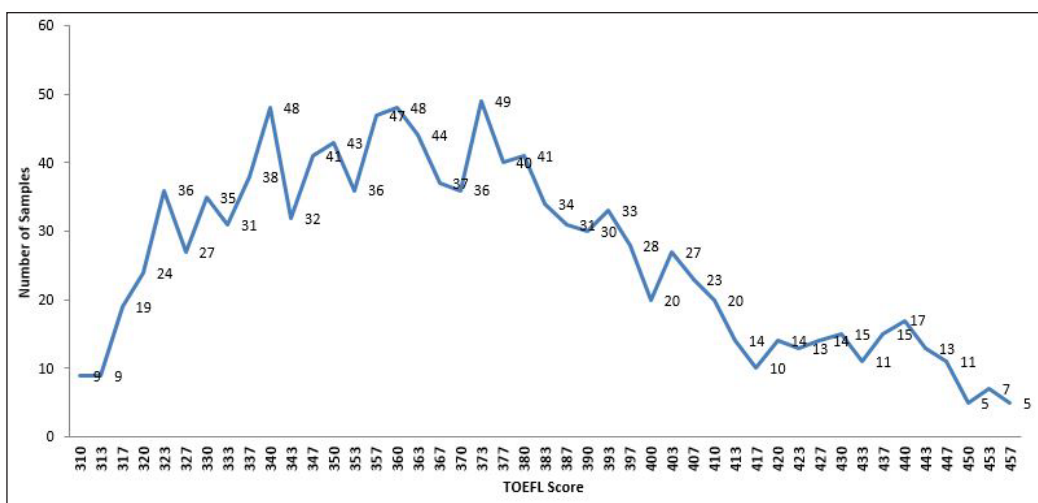

Figure 4. Number of samples evaluated for each score

Among those baseline scores evaluated, the lowest score that had p-values of One-Sample T-Test higher than 0.05 for both lower and upper deviations would be considered as the boundary where the rejection of null hypothesis, that the score had a deviation lower than or equal to 13 points, failed. This conclusion should be supported by the Shapiro-Wilk test, the normality assumption test, that should have p-values higher than 0.025. The One-Sample T-Test result is presented in Figure 5 below.
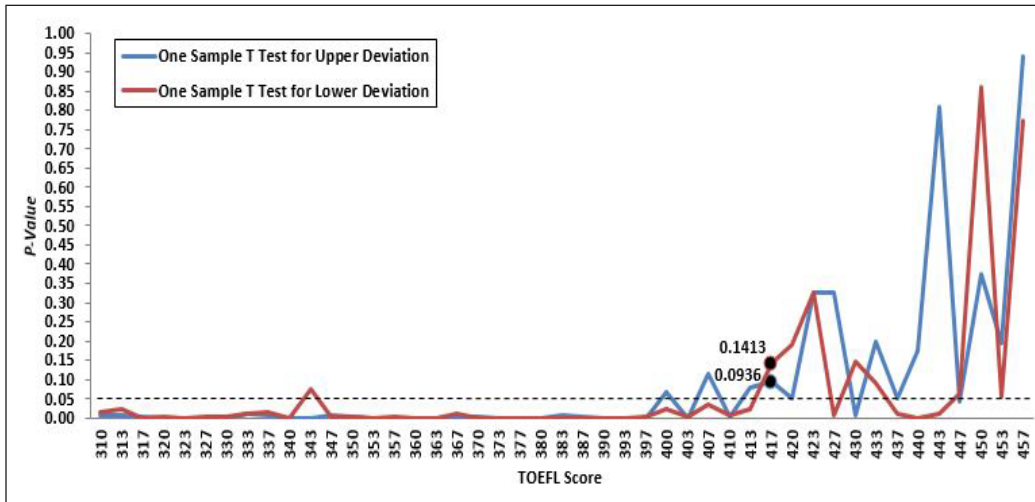

Figure 5. One-Sample T-Test

Figure 5 above shows that the lowest baseline score that failed to reject the null hypothesis was 417. Further, the One-Sample T-Test showed that the lower deviation at this score had a p-value of 0.094, while the upper deviation p-value was 0.141. Moreover, the number of samples at this baseline was ten scores, with p-values for the Shapiro- Wilk Test of 0.053 and 0.029 for the lower and upper deviation respectively. The detail is given in Table 4.

Table 4
Summary of the test for the score of 417

Variables		Statistics
TOEFL Score Evaluated (Baseline)		417
Number of Samples Evaluated		10
P-value for Normality Test	Lower Deviation	0.053
	Upper Deviation	0.029
P-value for One-Sample T-Test	Lower Deviation	0.094
	Upper Deviation	0.141
Sample size		1,180 (7.87%)
Population size		15,000

Although the lowest score with deviations within 13 points was 417, stability was indicated at the score of 400, and it appeared better at 407. However, only the upper deviation, the deviation between the second and the third tests, satisfied the Standard Error of Measurement of 13 points at these scores.

DISCUSSION

The objective of this study was to find out the lowest score in paper-based TOEFL which can be used for placement or to judge the level of English proficiency. We hypothesized that if the score fluctuated higher than the Standard Error of Measurement, the score cannot be used for the given purposes. Therefore, a statistical analysis was used to test repeat TOEFL scores between 310 and 677 to find out the lowest score where fluctuations were within the Standard Error of Measurement, i.e. 13 points. A total of 1,180 scores were analyzed to determine the interval of fluctuations between tests at intervals of less than two weeks. The research result shows that stability first appeared at a score of 400 but only for the subsequent test not the preceding test. The scores stopped fluctuating at 417 for both previous and next tests.

The data revealed that greater fluctuations occurred between the first and the second tests/baseline, particularly in the range between 400 and 413 and between 437 and 443. This finding is expected because test takers are unfamiliar with the test on their first attempt. Test takers could also be anxious when they take the test the first time, this anxiety decreases once they have experienced a similar test (Young, 1991, p. 434). In addition, first timers were also test-naive, trying their best to answer all questions because "they overestimated their likelihood of passing the exam" (Nijenkamp, Nieuwenstein, De Jong, & Lorist, 2016, p. 15). When they did not pass and took the second test, they might have applied test-wiseness strategy or guessed randomly, which they also did in the third test. In addition, after failing the first test, the students have been found to do some revision (McManus, 1992, p. 61) and therefore could master some basic rules of grammar and reading sub-skills, and strategies for listening such as focusing on the second speaker, prediction, etc. At the third test, where fluctuations were more stable, they might have read the same materials or tried more advanced rules and strategies but failed to understand them. In addition, fatigue and boredom presumably contributed to this stable score fluctuation (McManus, 1992, p. 61).

Although these research findings do not invalidate the use of TOEFL for language training or as an admission requirement, these findings suggest that scores below 417 cannot be confidently used to judge the English proficiency of the test takers. The figure of 417 is only 16 points ahead of scores which can be obtained through random guessing, and three points further from scores obtained through test-wiseness strategy. Should TOEFL scores be used for placement in language training, those students whose scores are below 417 should

be placed in one class. Alternatively, for placement TOEFL should be accompanied by an additional test, such as an interview. It is indeed not recommended to base placement merely on TOEFL scores (Brown, 1996, p. 283). In the case that other tests are not feasible and it is essential to establish another class level, the score of 400, which is close to the maximum score which can be achieved by using test-wiseness strategy according to research by Allan (1992), and Yang (2000), can be used to divide the levels with caution.

CONCLUSION

TOEFL scores are widely used to measure students' English proficiency for placement, however there is potential for misinterpreting the scores, which can result in misjudgment or misplacement. Random guesses and test-taking strategies are two contributors to such misinterpretation. However, this study has predicted the maximum scores obtained through random guesses and test taking strategies combined, and discovered that starting from 417, all factors other than English proficiency have been eliminated.

There are some limitations to the current research. While the interval between the first and second tests, as well as the second and third tests, was controlled, the exposure to learning could not be monitored. If all test takers included in the sample were prevented from preparing before the three tests, the result would be more accurate. Therefore, there is room for further, improved research in this area. In addition, the Standard Error of Measurement used in analyzing the data was

provided by ETS, where the sample used to analyze it did not include participants in this research. Future research is encouraged to use Standard Error of Measurement obtained from the same data used for the data analysis. Consequently, the result will be more representative. Finally, the raw scores were converted to TOEFL scores by using the conversion table provided by Phillips (2003, p. 258). Using the real conversion figures from ETS, to which the authors did not have access, will definitely improve the quality of the research.

REFERENCES

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, and C. Doe (Eds.), *Language Testing Reconsidered* (pp. 21-39). Ontario: University of Ottawa Press.

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 Ed.).* New York: 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.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge

certificate exams, IELTS, and TOEFL. *System, 28*(4), 523-539.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System,* 41(3), 663-676.

Douglas, D. (2010). *Understanding Language Testing.* London: Routledge.

Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement*, *5*(4), 321-325.

ETS. (2005). *TOEFL® iBT scores: Better information about the ability to communicate in an academic setting*. New Jersey: Educational Testing Service.

ETS (2010). TOEFL research. *TOEFL iBT Research Insight*, *1*(2). Retrieved April 10, 2017, from https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_slv2.pdf

ETS. (2011). *Test and score data summary for TOEFL® Internet-based and paper-based tests*. New Jersey: Educational Testing Service. Retrieved April 10, 2017, from https://www.ets.org/Media/Research/pdf/TOEFL-SUM-2010.pdf

ETS. (2011a). TOEFL program history. *TOEFL iBT Research Insight*, *1*(6). New Jersey: Educational Testing Service. Retrieved April 10, 2017, from https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf.

ETS (2016). *TOEFL ITP® reliability table*. Retrieved April 10, 2017, from https://www.ets.org/s/toefl_itp/ pdf/ toefl_itp_score.pdf

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

Gear, J., & Gear, R. (1996). *Cambridge preparation for the TOEFL test (2nd Ed.).* Cambridge: Cambridge University Press, Cambridge.

Hatch, E., & Lazaraton, A. (1991). T*he research manual: Research design and statistics for applied linguistics*. Boston: Heinle & Heinle Publishers.

Heffernan, N. (2006). Successful strategies: Test-taking strategies for the TOEFL. *The Journal of Asia TEFL*, *3*(1), 151–170.

Kashkouli, Z., & Barati, H. (2013). Type of test-taking strategies and task-based reading assessment: A case in Iranian EFL learners. *Procedia - Social and Behavioral Sciences*, *70*, 1580-1589.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291-308.

McManus, I. C. (1992). Does performance improve when candidates resit a postgraduate examination? *Medical Education*, *26*(2), 157–162.

Millman, J., & Bishop, C. H. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707- 726.

Mustafa, F. (2015). Using corpora to design a reliable test instrument for English proficiency assessment. In *Proceedings of the 62th TEFLIN International Conference* (pp. 344-352). Denpasar, Indonesia: Udayana University Press.

Mustafa, F., & Apriadi, H. (2016). DIY: Designing a reading test as reliable as a paper-based TOEFL designed by ETS. In *Proceedings of the 1st English Education International Conference* (pp. 402-407). *B*anda Aceh, Indonesia: Syiah Kuala University Press.

Nijenkamp, R., Nieuwenstein, M. R., De Jong, R., & Lorist, M. M. (2016). Do resit exams promote lower investments of study time? Theory and

data from a laboratory study. *PLoS ONE*, *11*(10), 1–19.

Nisbet, D. (2002). *Language learning strategies and English proficiency of Chinese university students* (Unpublished Doctoral thesis), Regent University, United States.

Phillips, D. (2003). *Longman Preparation Course for the TOEFL. New York:* Pearson Education.

Pyle, M. A., & Page, M. E. M. (1995). *Cliffs TOEFL Preparation Guide* (5th ed.). Lincoln, NE, United States: CliffsNotes Inc.

Reid, F. (1977). An alternative scoring formula for multiple-choice and true-false tests. *Journal of Educational Research, 70*(6), 335-339.

Rosenfeld, M., Oltman, P. K., & Sheppard, K. (2004). *Investigating the Validity of TOEFL: Criterion-related Strategies*. New Jersey: Educational Testing Service.

Sabarun. (2012). The students' scores on the different Institutional TOEFLs at the sixth English Department students of the Palangka Raya State Islamic College. *Educate*, *1*(2), 30-43.

Spolsky, B. (1990). Prehistory of TOEFL. *Language Testing*, *7*(1), 98–118.

Takagi, K. (2011). *Predicting academic success in a Japanese international university* (Doctoral thesis). Available from Electronic Thesis and Dissertation (ETD) at Temple University Libraries. (Record No. 81713)

Tannenbaum, R. J., & Baron, P. A. (2012). *Mapping the TOEFL® ITP Tests onto the Common European Framework of Reference*. New Jersey: Educational Testing Service.

Tavakoli, E., & Samian, S. H. (2014). Test-wiseness strategies in Paper-baseds and IBTs: The case of EFL test takers, who benefits more? *Procedia - Social and Behavioral Sciences, 98*, 1876-1884.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 560-620). Washington: American Council on Education.

Wells, C. S., & Wollack, J. A. (2003). *An instructor's Guide to Understanding Test Reliability.* Madison: University of Wisconsin.

Yang, P. (2000). *Effect of test-wiseness upon performance on the Test of English as a Foreign Language* (Doctoral thesis). Available from National Library of Canada database. (Record No. 26401474)

Young, D. J. (1991). Creating a low-anxiety classroom environment: What does language anxiety research suggest? *The Modern Language Journal, 75*(4), 426-437.

Wainer, H., & Lukhele, R. (1997). *How Reliable is the TOEFL Test?* New Jersey: Educational Testing Service.

Way, W. D., & Reese, C. M. (1991). *An investigation of the use of simplified IRT models for scaling and equating the TOEFL test*. New Jersey: Educational Testing Service.