An Analysis of Item Difficulties in PBT for the Graduate Candidates of IAIN Bukittinggi

Merry Prima Dewi
{merryprimadewi@gmail.com}

English Department Institut Agama Islam Negeri (IAIN) Bukittinggi, Indonesia

Abstract. this research aimed at finding out the item difficulties of the listening, structure and written expression, and reading comprehension in Paper Based TOEFL (PBT) conducted at IAIN Bukittinggi. The most problematic issue in IAIN Bukittinggi is the TOEFL score. The mean score for the graduate candidates for the bachelor degree and diploma was 357. Based on the passing grade set by the institution, the graduate candidates should get at least 400. This is a descriptive quantitative research with the 1035 graduate candidates as the population. The total sampling was used to determine the sample. Item difficulty index is used to determine the classification of each item in the PBT. The result of the research shows that all of the items in the PBT test were considered difficult and moderate, none of the item was considered easy. It means that the PBT is very difficult for the students. This is the reason why the graduate candidates could not achive the passing grade set by the institution

Keywords. item difficulty, PBT, TOEFL score

1 Introduction

The Test of English as a Foreign Language (TOEFL) is an examination that is administered by the Educational Testing Service (ETS) and is used to evaluate a nonnative English speaker's proficiency in the English language[1]. The graduate candidates of IAIN Bukittinggi have to take Paper Based TOEFL (PBT) as one of the requirements to graduate. The score gained from this test will show their English proficiency level. They have to be able to achieve the passing grade set by the institute. If they cannot pass the test with the minimum score, they need to repeat the test until they pass. As the requirements set by the institute, they Graduate candidates need to pass the test with 425 as the minimum score for the English Department students and post graduate program and 400 for other majors[2]. English Department students is expected to have higher achievement than the non-English major, which is the reason why they have higher standard than the non-English Department Students. For non-English Department students, they still need to get 400 as their minimum score. But they are all expected to get the score much higher than the passing grade.

Until the 10th graduation, PBT was held as the requirement to graduate from this campus. But for the next graduation, the graduate candidate must pass the PBT by the passing grade set by the university. It will not be the requirement to graduate, but it will be the requirement to enter thesis examination. When they cannot pass the score set by the institute, they need to repeat the PBT test until they reach at least the minimum score. The students need to pass this

BICED 2019, October 17-18, Bukittinggi, Indonesia

DOI 10.4108/eai.17-10-2019.2289740

PBT in the assumption that they will go directly to the working filed or higher education level. By having the PBT certificate, it will help them to measure their English proficiency level. If they have sufficient knowledge of English, it will help them in finding a job, because most of the company will ask the ability of other languages, and one of them is English. In addition, if the graduates would like to continue their education, they will also need this certificate as one of the requirement to register in the new educational level.By having those facts above, the graduate students need to have good English competence. In the learning process especially for Bachelor degree in any major, they got several subjects related to English competence. From several English subject offered, they got the opportunity to develop their English competence. After finishing all the subject offered for their Bachelor degree, their competence in English will be tested before they have their final thesis examination.

There are three sections of the PBT; they are 50 items of Listening Comprehension, 40 items for Structure and Written Expression, and 50 items of Reading Comprehension. Therefore, the total items for PBT is 140. Furthermore, the length of time to finish listening section is approximately 40 minutes, 25 minutes for structure and written expression, and 55 minutes for reading comprehension section. Thus, the amount of time needed for one test is approximately 120 minutes (2 hours)[3].The minimum score of the PBT is 217 (when none of the answer was correct) and the maximum score is 677 (when all the items were correct). It means that the students from the English Department need 208 to pass the test, while for non-English Department students only need 183. In fact, the students had an issue to achieve the passing grade. They still got low score. As a result the must repeat the test in order to pass the test and graduate from IAIN Bukittinggi. Based on the data from the Language Center of IAIN Bukittinggi (the unit which provides the PBT) the result of the PBT was unsatisfactory. As the sample, the data from the 8th graduation showed that the mean score was 357[4]. It means that the majority of the graduate candidates could not reach the passing grade. There must be something behind this issue. Most of the students could not answer the PBT items correctly, as a result, they got low PBT score. Based on the data above, the researcher was interested to conduct an analysis of item difficulty in PBT (Listening Comprehension. Structure and Written Expression, and Reading Comprehension section) for the candidate graduates of 8th graduation of IAIN Bukittinggi, In order to be precise, the meaning of the term test items should be clarified. There are some expert purposes the meaning of the term item and test item. First, Osterlind mentioned that an item is a particular element in a series or collection and is specified independently. In addition, the term *test item* is broad enough to allow for a variety of item formats and item classifying categories.[5]. The next expert is Haladyna who mentioned that a test item is a device for obtaining a answer, that is consequently scored using a scoring rule. All item formats have the same components: 1). A question or command to the test taker, 2). Some conditions governing the response, and 3). Scoring procedure.

In short, a test taker must understand what they will do about the test they are taking. Furthermore, he mentioned three fundamental types of item formats:

1. *Objective versus subjective scoring*. The former type of scoring is clerical and with a very small degree of error. The latter requires a human judge using a descriptive rating scale. This kind of judgment usually contains a higher degree of random error.
2. *Selection versus production*. With the SR format, the test taker recognizes and selects the answer. With the CR format, the test taker produces the response. Producing a response usually implies higher fidelity to the task in the target domain.
3. *Fixed-response versus free-response*. Some CR items are written to offer the test taker more freedom to express oneself, whereas other CR items are more focused on generating a structured response.

4. *Product versus performance*. Some CR items have a product evaluated. Usually the product is a written document but it could be a model, invention, or another similar palpable object. A performance can be analyzed for certain qualities related to a predetermined process. The interest in performance is one of technique versus an outcome.

Two most common item characteristics are *difficulty* and *discrimination*. The most fundamental measure of item difficulty is the proportion (or percentage) of test takers responding correctly. This statistic is known as the item p-value. Every item has a natural difficulty. This value is based on the performance of all persons we intend to test. This p-value is very difficult to estimate accurately unless a very representative group of test takers is being tested.[6]. Kurpius mentioned that Item discrimination is the "the degree to which an item differentiates correctly among test takers in the behavior that the test is designed to measure" Some potential criterion groups consist of those who succeed or fail in an academic course, in a training program, or in a job. As a budding measurement specialist, you need to be able to pick tests that have strong item discrimination. [7]. The purposes of item analysis are (a) to evaluate the item during its initial tryout, (b) to verify the key—correct answer, and (c) to correctly estimate difficulty and discrimination for scaling for comparability, especially if multiple test forms are used. Knowing an item's difficulty and discrimination is a great help in test design.[6]

Kurpius stated that item difficulty shows proportion of person who got an item correct. Similar to a percentage score, item difficulty (p) might range from 0 to 1.00. The higher the proportion of people who get the item correct, the higher the value of *p.* [7] In this research, the researcher will focus on finding the of difficulty index of each item on the PBT. Based on the purposes suggested by Haladyna above, it is necessary to find out further about item difficulties. According to Domino item difficulties represents a scale of measurement identical with percentage, where the average is 50%and the range goes from zero to 100%. This is of course an ordinal scale and is of limited value because statistically not much can be done with ordinal measurement.

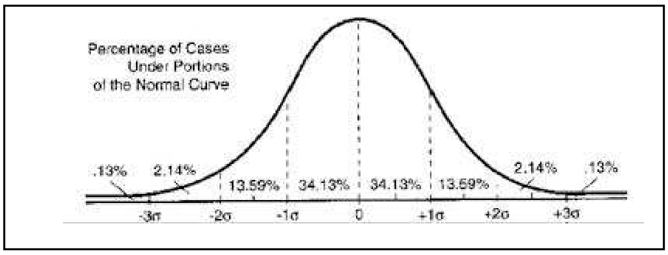

Fig.1. Percentage of cases under portions of the normal curve

This is the normal curve concerning items difficulty which need to be spread out evenly using the normal curve. To be brief, there are small amount of easy items and difficult items, and mostly moderate items.[8] Furthermore, Osterlind mentioned that discrimination is another important concept for judging the quality of items. Actually, we were examining discrimination for items in the preceding section, but it may not have been conceptually understood. Discrimination for items may be conceptually understood as the relationship between the difficulty of an item and the ability of the examinees. Simply put, item discrimination is an index for determining differences among individual examinees on the subject matter or psychological construct being assessed.[5] In addition Domino stated that Item discrimination refers to the ability of an item to correctly “discriminate” between those who are higher on the variable in question and those who are lower.

Moreover, Wilson stated that item discrimination is related to item difficulty because it is an indicator of how well the item discriminates between students’ who know the material and students who don’t. Item discrimination is determined using students’ total test scores as an indicator of their competence with the material being assessed. In other words, in general, high scoring students would be expected to answer items correctly, and low scoring students would be expected to answer items incorrectly. When all of the low scoring students answer an item correctly, and the high scoring students answer it incorrectly, the item did not function as expected.[7]. Concerning about the theories above, the research focused on the finding the item difficulty index for each items on PBT. Item discrimination would not be covered in this research since the researcher did not mean to discriminate students based on their PBT test, but focusing on finding out the problems behind the low scores gained by the gradute candidates of IAIN Bukittinggi.

2 Research Method

This is descriptive quantitative research which aimed at finding out the level of difficulty level of the three sections of PBT. There were 1035 graduate candidates as the population. They were the graduate candidates for the 8th graduation of IAIN Bukittinggi, and they took the PBT test. Total sampling was used to find out the exact number of students who faced problems in each items and each sections of the test. The PBT test score were analyzed per section, beginning from the 50 items of Listening Comprehension, 40 items of Structure and Written Expression, and 50 items for Reading Comprehension. The after each item on a PBT was corrected, the number of correct answer and incorrect answer per items were analyzed. The researcher used the following formula to find out the level of item difficulty for each items. The index of item difficulty *(p):*

$$p = \frac{\textit{the number of testee who answer the item correctly}}{\textit{number of testee}}$$

Table 1. Level of Difficulty Index [9]

Range	Difficulty level
< 0.3	easy
0.3 – 0.7	moderate
> 0.7	difficult

The result of the analysis of item difficulty for the Listening comprehension section were as follows:

Table 2. Level of difficulty for listening comprehension section

Level of item difficulty	The number of items	Percentage
easy	0 item	0%
moderate	22 items	44%
difficult	28 items	56%

From the table above, it was shown that none of the Listening Section items were considered easy items. The majority of the items were difficult. More than a half of the items (28 items) in listening section were considered difficult. It means that 56 % of the listening items is too difficult for the students to answer, so that no one can answer the items correctly. The moderate level items for Listening section were 22 items which classified as the moderate items. It means that 44% of the items were in moderate level. In fact, there were no items in the Listening Comprehension was considered easy for the students to answer. Based on the result of the Listening Comprehension section taken by 1035 students of IAIN Bukittinggi, the Listening comprehension section is considered a hard section since 56% of the testee could not answer them correctly.

The result of the analysis of item difficulty for the Structure and Written Expression section were as follows:

Table 3. Level of difficulty for structure and written expression section

Level of item difficulty	The number of items	Percentage
easy	0 item	0%
moderate	20 items	50%
difficult	20 items	50%

From the table above, it was shown that none of the Structure and Written Expression items were considered easy items. The majority of the items were difficult. A half of the items (20 items) in Structure and Written Expression section were considered difficult. It means that 50 % of the listening items is too difficult for the students to answer, so that no one can answer the items correctly. The moderate level items for Structure and Written Expression section were 20 items which classified as the moderate items. It means that 50% of the items were in moderate level. In fact, there were no items in the Structure and Written Expression section was considered easy for the students to answer. Based on the result of the Listening Comprehension section taken by 1035 students of IAIN Bukittinggi, the Structure and Written Expression section is considered a hard section since 50% of the testee could not answer them correctly.

The result of the analysis of item difficulty for the Raeding Comprehension section were as follows:

Table 4. Level of difficulty for reading comprehension section

Level of item difficulty	The number of items	Percentage
easy	0 item	0%
moderate	23 items	46%
difficult	27 items	54%

From the table above, it was shown that none of the Reading Comprehension items were considered easy items. The majority of the items were difficult. More than a half of the items (27 items) in Reading Comprehension section were considered difficult. It means that 54 % of the Reading Comprehension section was too difficult for the students to answer, so that no one can answer the items correctly. The moderate level items for Reading Comprehension section were 23 items which classified as the moderate items. It means that 46% of the items were in moderate level. In fact, there were no items in the Reading Comprehension section was considered easy for the students to answer. Based on the result of the Reading Comprehension section taken by 1035 students of IAIN Bukittinggi, the Reading Comprehension section is considered a hard section since 50% of the testee could not answer them correctly.

The normal curve of the difficulty index was not found in this PBT test. Because the level of the difficulty only concentrated in the difficult and moderate level. From the three sections; Listening Comprehension, Structure and Written Expression, and Reading Comprehension, there was no item which was considered easy. In short, from 140 PBT test items, they were all considered moderate and difficult for the testee.

Table 5. Level of difficult for PBT test for 140 items

Level of item difficulty	The number of items	Percentage
easy	0 item	0%
moderate	65 items	46.42%
difficult	75 items	53.57%

From the table above it can be seen that the PBT test conducted by Language Center of IAIN Bukittinggi for the 8th graduation by having 1035 testee was considered very difficult. It was the reason that most of the students could not get the passing grade set by the institute. Most of the students only got 357 as their score. It was because of the items on the PBT were very difficult for them to answer. The table shows that from the 140 items of PBT, 53.57 % was a difficult level, while 46.42% was considered moderate items, and no items from was considered an easy item. Regarding to the analysis of the data gained from the 8th graduation graduate candidates, it is suggested for the Language Center of IAIN Bukittinggi to review PBT test items. The items which was considered difficult should be revised, some of the moderate item can still be used, and it is suggested to add easy level items for the testee.

1 Introduction

[1] J. G. and R. G. Gear, *Cambridge Preparation for the TOEFL Test 3rd Edition*. Cambridge: Cambridge University Press, 2006.

[2] Rector IAIN Bukittinggi, “SK TOAFL DAN TOEFL SYARAT.pdf,” Bukittinggi, 2019.

[3] A. E. Estiwi Reto Purnaning, Alvina Kusuma Ayuningtyas, Nurul Huda, *Big Book TOEFL*. Jakarta: KawahMedia, 2014.

[4] U. P. B. I. Bukittinggi, “No Title,” Bukittinggi, 2018.

[5] S. Osterlind, *Constructing Test Items : Editors : Other books in the series :* New York: Kluwer Academic Publishers, 2002.

[6] T. M. dan M. C. R. Haladyna, *Developing and validating test items*. New York: Routledge, 2013.

[7] L. W. Wilson, *What Every teacher needs to know about assessment*. New York: Routledge, 2005.

[8] M. Windows, M. Corporation, K. Hori, and A. Sakajiri, *Psychological Testing An Introduction*. New York: Cambridge University Press, 2006.

[9] R. Utley, *Theory and Research for Academic Nurse Educators: Application to Practice*. Ontario: Kevin Sullivan, 2011.