REVISITING CULTURAL BIASES IN TOEFL TEST: A PILOT STUDY

Nyak Mutia Ismail1*, Arisna Mukti1, Marisa Yoestara2, Hilda Delavari3 & Alireza Qadiri3

1Department of English Education, Faculty of Teacher Training and Education, Universitas Syiah Kuala, Banda Aceh, Indonesia
2Department of English Education, Faculty of Teacher Training and Education, Universitas Serambi Mekkah, Banda Aceh, Indonesia
3Mehryar Institution of Higher Education, Yazd, Iran

*Corresponding author: nyakmutiaismail2010@gmail.com

Abstract

TOEFL test is required as one of the standardized test in English proficiency for scholars worldwide, across cultures. This study is a pilot study conducted to see cultural bias contained in the TOEFL test in various TOEFL Preparation modules that are frequently used as the source of questions for TOEFL equivalent in Aceh. Descriptive content analysis was considered as a method that well-conforms the needs of this study. The data were taken from four different TOEFL modules—Cambridge TOEFL Preparation, Barron’s TOEFL Practice, Cliff’s TOEFL Practice, and Longman TOEFL practice module.—specifically scoped into reading section alone. Hence, the instruments involved in this study were the documents of TOEFL preparation module. The questions which indicate cultural bias are verified as the data of this study. Later on, the data analysis was carried out by employing three-step analysis by Miles, Huberman, and Saldana (2014). The result shows that from those four TOEFL modules, particularly in reading section, the quantities of biased items are not cautious for EFL TOEFL test-takers. However, future research covering the Listening and Structure and Written Expression sections should be taken into account.

Keywords: Standardized test, TOEFL test, constructed bias, cultural bias, test bias.

INTRODUCTION

Test of English as Foreign Language (TOEFL) is a test administered by the Educational Testing Service (ETS) and used to assess non-native English speakers in order to find out ones’ English language proficiency level and to be used on the specific requirements of most institutions or agencies (Gear & Gear, 2002). Evaluating language proficiency is undoubtedly important as stated by Pan (2014) that many Asian countries now require the students and the citizen to have a certain level or score on English language proficiency in order to graduate and to get a job in order to improve their English aptitude and their global market competitiveness. However, it has become a sensitive issue as the exams where English is used as the second language or the foreign language like in most Asian countries are usually offered at limited times and expensive (Sims, 2015).

Another issue is that the results of English language proficiency test taken by the students can exhibit great variety since the level of the students learning English varies significantly around the world, and those various forms of English language include Australian, Canadian, New Zealander, Irish, British and American (Adedayo, 2015). It is not an easy task for the students to reach a certain level or score on the exams since many of them are not familiar with the material examined. To solve the problem, universities have to help the students who did not get the required score by providing support/alternative/complementary measures (Pan, 2014) so that they can fulfil the language proficiency requirement.

Testing processes are often plagued by serious sources bias as the language proficiency test such as TOEFL is a standardized test usually designed to determine whether the test takers are ready to study in a United States or Canadian university (Pyle, 2001). As classified by Van de Vijver and Poortinga (1997) in Van de Vijver and Tanzer (2004), there are three most typical sources of bias in the test due to the cross-cultural differences. First, it is construct bias. Potential sources for this type of bias is that the passage only elaborates partially the definition of the construct across cultures, differential appropriate of the behaviors associated with the construct (e.g., skills do not belong to the repertoire of the cultural
groups), poor sampling of all relevant behavior (e.g., short instruments), and incomplete coverage of all relevant aspects/facets of the construct (e.g., not all relevant domains are sample). Second, it is method bias. The sources are from the incompatibility of sample (e.g., caused by differences in education, motivation), differences in environmental administration conditions, physical (e.g., recording devices) or social (e.g., class size); ambiguous instructions for respondents and/or guidelines for administrators, differential expertise of administrators, tester/interview/observer effects (e.g., halo effects), communication problems between respondent and tester/interviewer (including interpreter problems and taboo topics), differential familiarity with stimulus material, differential familiarity with response procedures, and differential response styles (e.g., social desirability, extremity scoring, acquiescence). Last, it is item bias from which it can root from the poor item translation and/or ambiguous items, nuisance factors (e.g., item may invoke additional traits or abilities), and cultural specifics (e.g., incidental differences in connotative meaning and/or appropriateness of the item content).

Van de Vijver and Tanzer (2004) further explained that “sample bias or incomparability of samples occurs when the samples used differ in a variety of relevant characteristics (‘nuisance factors’) other than the target construct. Administration bias includes all sources of bias which are caused by the particular form of administration. Instrument bias subsumes all sources of method bias which are associated with the particular assessment instrument”. The sources of bias namely “raters, items, tests as well as gender, age, race, language background, culture, and socio-economic status need to be spotted and removed” (Shirazi, 2019) as they can lead to test bias that might cause the score to be inaccurate (Xi, 2010).

On the other hand, Angoff (1989) carried out a study regarding whether or not the item refers to American history, geography, and so forth would benefit the examinees who have known about American culture tested in the TOEFL by giving the same test to a group of students who had lived in the United States for more than a year and a group of students in non-English speaking countries who had never lived in the United States more than a month. The results of the study may be concluded that it did not affect the test result. Furthermore, Reynolds and Suzuki (2003) stated that the controversy of bias in standardized test is stirred largely by public concerns as they often conjure up notions of prejudicial attitudes. The statement is also supported by Chen and Henning (2015) that the long-standing concern over test bias partly influences the nature of psychological characteristics and their measurement. Moreover, research indicates that a source of measurement bias is actually a prejudiced of stereotype that it is unfair to test students’ comprehension about other cultures (Wichert, Dolan & Hessen, 2005). In fact, it is completely different between the concepts of test bias and unfairness, but it is unarguable that “a test may have very little bias, but a clinician could still use it unfairly to minority examinees’ disadvantage” (Reynolds & Suzuki, 2003).

METHODS
The method used in this study was descriptive content analysis. The method works well on studying a nature of text, situation, images, audio, or video without having to intervene it with questionnaires, interview, or other stimulating experiences (Bachman, 1990). This study scoped to examining texts only. Concerning its validity, the validity of this study is controlled by the consistency of the analysis (Weber, 1990). Regarding the consistency issue, there were four TOEFL modules chosen to be analyzed in this study; these include Cambridge TOEFL Preparation, Barron’s TOEFL Practice, Cliff’s TOEFL Practice, and Longman TOEFL practice module. To investigate the existence of culturally-biased items in various types of TOEFL test, we used practice books for different types of the test. In this regards, for all the modules, PBT practice books were firstly analyzed. Then, we examined Cambridge CBT preparation book and Barron’s IBT book as well. In each module, only the complete tests were examined for their reading section. Thus, the reading passages along with the following questions were carefully read to be considered whether the question item has cultural bias or not. The bias criteria are as suggested by Van de Vijver and Poortinga (1997); they are that the passage/stem item only elaborates partial the definition of the cultures, lack of mutual understanding on the behaviors associated with the item, lack of examples given to the relevant behavior, and incomplete description on the relevant item. Later on, the data were analyzed using three-step analysis where the irrelevant data were reduced in the data reduction phase; next, the data were displayed; and finally the data were verified using various theories and postulations from other experts and researchers.

RESULTS AND DISCUSSION
The following narration conceives the cultural biased items in each module examined. First, the Longman TOEFL module was examined. It was found that there are three items that contain a cultural bias. It is item number 37 on page 432 and item number 42 on page 459. The items are as below. The final item is number 38 on page 484.

The expression "weathered the storm" or in line 23 could best be replaced by
(A) found a cure for
(B) rained on the parade of
(C) survived the ordeal of
(D) blew its stack at

A “heyday” in line 1 is most probably a
(A) time for harvest
(B) a period with low prices
(C) a period of great success
(D) a type of railroad schedule

Which of the followings would most likely have taken place at Richard’s Madison Square Garden?
A ballroom dance
A theater production
A basketball game
A tiger show

The first two questions above are not easy to answer if the test-takers do not have any information about idioms used in American culture. Since idioms are strongly interrelated with culture, students cannot even employ any guessing strategy to identify the correct answer. Furthermore, the expressions ‘weathered the storm’ and ‘heyday’ indicate to a certain weather condition in a certain place, particularly United States of America. Thus, EFL students in Indonesia are completely unfamiliar with American climate and weather. Blau, Moller, and Jones (2004) conjointly support that where students experience life profoundly predisposes their success in answering the questions in a high-stake, standardized test. For example, an Acehnese student who grew up in Aceh would certainly acquire Acehnese geographical condition, climate, racial issues, and socio-economic issues of Aceh. Thus, a certain question set about Aceh would most probably be effortless for him/her. In respect with the third question, the test-takers may face difficulties if they are not familiar with the cultural-historical background behind Richard’s Madison Square. Clearly, not everyone has sufficient knowledge about what has historically taken place there.

Later, after reexamining Barron’s TOEFL module for Paper-based Test, the following culturally biased items are found as in the following. There is only one item which is found to be culturally biased in this module; it is item number 18 on page 324:

Along the San Andreas Fault, tremors are
(A) small and insignificant
(B) rare, but disastrous
(C) frequent events
(D) very unpredictable

The question above requires a student to understand the geography of America—which is almost never learned by EFL students beyond the Americas. This item is also considered as a culturally biased item for that reason. Indeed, the passage has provided information about it. However, the information does not meet the need of the EFL test-takers sufficiently. The test-takers need to have a grounded knowledge about the San Andreas Fault, where it is located, why it is always perplexed with tremors and earthquakes, and so on. Vague knowledge cannot help test-takers to decide the best answer (Djiwandono, 2006).

Moreover, three culturally-biased items were also found in Barron’s TOEFL module for the Internet-based Test. These include item 4 on page 346 and item 11 on page 347.

Why does the author mention the Forbes 400 in paragraph 3?
To explain the meaning of the listing that appears every year
To support the statement that most wealthy people inherit their money
To cast doubt on the claim that family income predicts individual wealth
To give examples of successful people who have modest family connections

All of the following are indicators of prestige in the United States, EXCEPT
The level of education that a person has achieved
The amount of money that an individual has acquired
The type of employment that someone pursues
The hard work that a person does on a consistent basis

As it can clearly be seen, the above questions are completely linked to American culture. The first item is connected with a list annually published in the U.S, familiarity with which can facilitate answering this question. It is obvious that answering this question requires some degree of familiarity with American culture. The second item deals with indicators of prestige in the U.S, which are clearly linked to the specific culture in the U.S and seems easier to test-takers who are well familiar with their culture.

Next, the Cliff’s TOEFL module was also examined and the result is as follows. Item number 37 on page 263.

Which of the following was not mentioned as a drawback of the conventional method of boning?
   a. storage space requirement
   b. energy waste
   c. loss of carcass weight
   d. toughness of meat

   Hot boning is one of advanced technique applied in modern/developed country in meat processing. This inference question is hard to be answered if the candidate does not have enough background of knowledge about it. The high level of vocabulary used makes it even worse. The technological improvements that take place in developed countries are not likely to be understood by people in developing countries such as Indonesia. Omi and Winant (2015) see that high-stake, standardized tests are racial projects where there is a construct of a chain of questions about an actual group of people (as the reference group) to be answered and comprehended by another completely different group of people (as the focal group), sharing none of cultural norms and beliefs.

   Then, the bias is also found in item number 43 on page 264.

In approximately what year will a different government take control of the Panama Canal
   a. 2000
   b. 2100
   c. 3100
   d. 2999

   The text did mentioned the taking over of the canal will be in the end of the 20th century. Candidate will be confused whether the answer is 2000 or 2100, whether 2100 is still included in 20th century or it is already enter into 21st century.

   Later, item number 56 on page 373 and item number 56 on 374,

Which of the following could not become president of the United States?
   a. a 38-year-old man born in the United States
   b. a 40-year-old naturalized citizen who has lived in the United States for 25 years
   c. a 45-year-old woman born of parents who were U. S citizens
   d. a 60-year-old woman who was born in Japan of parents who were citizens of the United States and who has lived in the United States for 20 years

   What is the maximum number of years that a person can serve as president of the United States?
   a. 4
   b. 8
   c. 10.
To answer these questions, the test-takers need to understand the concept of naturalized born and background of knowledge about U.S constitution inference question. As mentioned earlier, in a glance, the history and constitutional form of international countries are apparently learned in Indonesian school curriculum. However, the content learned only a little and it is insufficient to answer such questions in TOEFL test. This means that EFL test-takers in Indonesia should be provided with more elaborated lessons about US constitution in the TOEFL preparation classes so that what is learned in the classroom come to be questioned in the test items. In short, to offer the cultural questions in the test items, cultural relevant curriculum should be provided during classroom time (Young, 2010).

Finally, from Cambridge TOEFL Preparation module for Paper-based test, there is only one culturally biased item found. It is item number 45 on page 399. The item is as follows.

**According to the passage, before World War II...**

- Educational testing was widespread
- The military elite was better educated
- Individual opportunity had been increased
- Selection and individual opportunity were unconnected

The item above is seen as a culturally biased item because the test candidates have to recall the information about what happen in the US during World War II. Indeed, history subject over the world includes the topic about World War II, but not strictly detail explanations about the US alone. In the passage, the information about US Navy and test-taking agenda is mentioned but there are not enough samples to support this construct. Test-takers need the content knowledge about USA during World War II to decide the answer precisely. Van de Vijver and Tanzer (2004) have clarified that items with overlapping constructs across cultures without providing enough elucidating samples are seen as construct-biased items.

This is while no culturally-biased item was found in Cambridge TOEFL module for the Computer-based test.

Only to mention, when discussing about culture, we can see two contrasting culture between the reference groups (American) and the focal group (Indonesian/Asian) if the TOEFL test takers are Indonesian students. As Kim and Zabelina (2015) argue that as the TOEFL team constructs the TOEFL test items, their culturally specific standpoints get attached and embodied each item, e.g. self-image, the utilization of structure, and so on. This condition also leads to bias for EFL TOEFL test-takers. For example, in Bahasa Indonesia, the correct form to ask name is ‘Siapa namamu?’ which mistakenly appears to be ‘Who is your name? [sic]’ in English. In addition, studies also suggest that self-perception exhibits bias in standardized tests. American people—also European and Canadian—have propensities to self-enhancing, while Asian students tend to me more self-critical. So that, Asian students are more susceptible toward failure than success but their counterparts are more sensitive to success than failure (Heine, et.al, 2001). This self-image also weighs vital consideration in TOEFL accomplishment.

**CONCLUSION**

From the results, it can be concluded that—indeed—there are culturally biased items in TOEFL Reading Section despite the numbers of such items are not alarming. From Cambridge, there is only 1 item; Cliff, there are 4 items; Barron, there are 4 items; and Longman, there are 3 items. However, informing the test-takers about content knowledge on US culture such as geography, history, constitution, arts, and so on would not hurt. Students can even advantage from such knowledge. What appears as the limitation of this study is that it only examined the Reading section of TOEFL. In this section, the majority of topics are about science and very little found on other discussions. It is gently suggested that future studies would reexamine the Listening section in which—probably—most of the culturally biased items can be found. Moreover, future researchers are recommended to conduct similar studies on other modules such as NTC’s Preparation for the TOEFL, also on newer versions of the modules examined in this work. Additionally, various types of TOEFL including CBT, PBT and IBT can specifically be investigated.
REFERENCES


