ISSN: 1994-4217 (Print) 2518-5586(online)

Journal of College of Education

Available online at: https://eduj.uowasit.edu.iq

Asst. Lect. Jiyar Othman Hamadamin

College of Languages, Salahaddin University-Erbil

Email:
zeyar.hamadamin1@su.edu.krd

Keywords:

TOEFL iBT, Validity, Topic Effect

Article info

Article history:

Received 1.Jul.2024

Accepted 29.Jul.2024

Published 25.Aug.2024

The Validity of TOEFL iBT Reading Section: Reassessment and Evaluation

A B S T R A C T

The TOEFL iBT, which is a reputable test designed to measure such skills as listening, reading, speaking, and writing, has, for over a decade, caused a heated debate among scholars This study examines the reading section of this test to determine if it adequately covers the materials that students encounter in an academic setting. Furthermore, it also aims to investigate whether the reading section properly assesses what it is designed to gauge or if it involves other irrelevant constructs, which might consequently overshadow the individual differences. The study analyzed 209 authentic passages drawn from five different books published in different years to evaluate the content and construct validity as reliable indicators of reading proficiency. The categorization of the passages was based on the classifications put forward by Collins and Sorrenson (2014). Also, the word frequency of both the passages and the vocabulary questions were classified using Collins Dictionary's (2023) categorization of word frequency to determine whether or not the reading passages include many low-frequency words. Additionally, the study investigates multiple-choice questions (MCQs) using Haladyna's (2022) guidelines to see if they accurately assess reading comprehension without posing a threat to test validity. Although the study discovered that this test leaned toward hard science subjects, which may obscure the true and diverse intellectual abilities among testees, it also found positive aspects, such as limited low-frequency words and well-designed MCQs. The study, thus, recommends that TOEFL iBT test designers incorporate a wide range of topics from different fields to better represent a diverse student population.

DOI: https://doi.org/10.31185/eduj.Vol56.Iss2.3977

صلاحية قسم القراءة في اختبار التوفل عبر الأنترنيت: مراجعة وتقييم

م.م. ژيار عثمان حمد امين
كلية اللغات / جامعة صلاح الدين – أربيل

الملخص

كان اختبار TOEFL iBT، وهو اختبار ذائع الصيت بين الأوساط الأكاديمية، مصمم لقياس مهارات مثل الاستماع والقراءة والتحدث والكتابة، لأكثر من عقد من الزمان مثار جدل حاد بين الدارسين في مجال اللغة. تتناول هذه الدراسة قسم القراءة في هذا الاختبار لتحديد ما إذا كان يغطي بشكل مناسب المواد التي يواجهها الطلاب في البيئة الأكاديمية. علاوة على ذلك، يهدف أيضًا إلى التحقق مما إذا كان قسم القراءة يقيّم بشكل صحيح ما تم تصميمه لقياسه أو إذا كان يتضمن بنيات أخرى غير ذات صلة، والتي قد تطغى بالتالي على الفروق الفردية. قامت الدراسة بتحليل ٢٠٩ فقرات أصلية مستمدة من خمسة كتب مختلفة نشرت في سنوات مختلفة لتقييم المحتوى وبناء الصلاحية كمؤشرات موثوقة لكفاءة القراءة. تم تصنيف المقاطع بناءً على التصنيفات التي طرحها كولينز وسورنسون (٢٠١٤). أيضًا، تم تصنيف تكرار الكلمات لكل من المقاطع وأسئلة المفردات باستخدام تصنيف قاموس كولينز (٢٠٢٣) لتكرار الكلمات لتحديد ما إذا كانت فقرات القراءة تتضمن العديد من الكلمات الأقل تكراراً أم لا. بالإضافة إلى ذلك، تبحث الدراسة في أسئلة الاختيار من متعدد (MCQs) باستخدام إرشادات هالادينا (٢٠٢٢) لمعرفة ما إذا كانت تقيّم فهم القراءة بدقة دون أن تشكل تهديدًا لاختبار الصلاحية. وعلى الرغم من أن الدراسة اكتشفت أن هذا الاختبار كان يميل نحو المواد العلمية الصعبة، الأمر الذي قد يحجب القدرات الفكرية الحقيقية والمتنوعة لدى المشاركين، إلا أنها وجدت أيضًا الجوانب الإيجابية، مثل الكلمات المحدودة الأقل تكراراً وأسئلة الاختيار من متعدد الأسئلة المصممة بشكل جيد. وبالتالي، توصي الدراسة بأن يقوم مصممو اختبار TOEFL iBT بدمج مجموعة واسعة من المواضيع في مجالات مختلفة لتمثيل مجموعة متنوعة من الطلاب بشكل أفضل.

الكلمات المفتاحية: اختبار توفل عبر الانترنيت ، الصلاحية ، فعالية الموضوع

1. Introduction

The TOEFL test, which is created and administered by a non-profit organization known as Educational Testing Service (henceforth ETS), is a reputable test designed to assess and demonstrate students' English language ability (Alderson, 2009; Harsch et al., 2017; Li, 2018). This test was first designed in 1964 to use it as proof of English proficiency for those foreign students who, at that time, had the plan to enrol in those universities where English was the medium of instruction (Alderson, 2009; Harsch et al., 2017; Li, 2018). The TOEFL iBT, which was introduced in 2005, is now approved for admission by over 11,500 universities and institutes in more than 160 countries, including New Zealand, Australia, Canada, and the United Kingdom (Axe et al., 2020; Educational Testing Service, 2023; Harsch et al., 2017; Li, 2018; Toker, 2019).

The notion of TOEFL iBT test, reading section in particular, and its accurate measuring of examinees' English level has been at the forefront of discussion over the past decades. ETS and independent researchers have taken a keen interest in exploring the validity and reliability of this test and have published a wealth of research in this regard (Axe et al.,

2020). The adherents of this test acknowledge that this test does measure test takers' English level accurately; the topics included, however, are representative of what test takers might face in universities (Axe et al., 2020). One merit that can be ascribed to this test, some scholars note, is that test takers' prior knowledge to understanding the reading passages is not necessary and, thus, marginal because all the information needed to answer the questions can be easily deduced from the passages (Cohen & Upton, 2006; Collins & Sorrenson, 2014; Liu et al., 2009; ETS, 2017; Axe et al., 2020; Phillips, 2015).

On the other hand, this proposal, nonetheless, received backlash from critics who claim that this test, albeit designed by professional test designers, may not accurately gauge test takers' ability in English as the topics and the questions are designed in a way that requires testees to hinge partly, if not mostly, on prior knowledge (Hammad, 2021; Liu, 2011; Ovilia, 2018; Toker, 2019).

The studies conducted on the validity and reliability of the TOEFL iBT test mostly revolve around the issue of whether or not there is a correlation between test takers' academic achievements and their TOEFL scores (Cho & Bridgeman, 2012; Feast, 2002; Fox, 2004; Golder et al., 2009; Hill et al., 1999; Lee & Greene, 2007; Manganello, 2011; Vu &Vu, 2013; Wait & Gressel, 2009; Woodrow, 2006). However, few studies have been conducted on delving into the issues pertinent to the content and construct validity of the reading section of this test in a single study. Thus, this paper aims at critically evaluating whether or not the reading section of this test is representative of all the types of reading topics test takers face in real academic life. Its aim is also to investigates the extent to which the reading section (including its MCQs) of this test gauges test takers' ability to utilize and understand academic English topics. In other words, the reading passages will come under scrutiny to determine how well ETS's measurement method targets their claim that the reading section measures the extent to which one can "read and understand the kinds of materials used in an academic environment" (Educational Testing Service, 2022, para. 1).

Moreover, TOEFL critics claim that the reading section of this test covers scores of technical and low-frequency words and ignores a range of factors about individual differences, including test takers' preferred learning style, gender, background knowledge, culture, personalities, cognitive aptitudes, and physical characteristics (Ahmadjavaheri & Zeraatpishe, 2020; Bachman, 1990; Bachman et al., 1995; Bachman & Palmer, 1996; Fahim et al., 2010). Low-frequency words in reading comprehension passages and the aforesaid individual differences, as critics of standardized tests contend, tend to have a negative impact on test takers' comprehension of the subject matter (Panahi, 2014). Low frequency words, as opposed to high-frequency words, refer to those words that are used less commonly or frequently. For this reason, the study aims to both investigate whether or not the TOEFL iBT reading section includes too many technical and infrequent vocabulary items and the extent to which technical and low-frequency words, if any, in the TOELE iBT reading passages (including in the vocabulary questions) have a negative impact on test takers' performance in the test. Further, it also attempts to explore the extent to which these individual differences may or may not influence students' achievement in the test.

2. Literature Review

2.1. Issues Surrounding the Validity of the TOEFL iBT Test

The issue of whether or not the TOEFL iBT test is valid and a predictor of academic success has, for a decade, caused a heated debate among scholars. It is maintained that the content of the test corresponds to what test takers might face in a genuine academic setting and that test designers do not include inauthentic materials in an endeavor to both substantiate the validity of the test and refrain from designing an inaccurate and biased test that measures irrelevant constructs (Axe et al., 2020; Cohen & Upton, 2006; Esfandiari et al., 2018; Collins & Sorrenson, 2014; Liu et al., 2009; Phillips, 2015). Others assert that the materials included in this test do not require test takers to hinge heavily on prior knowledge, as all the information required to understand the materials and answer the questions can easily be gleaned from the texts provided (Axe et al., 2020; Cohen & Upton, 2006; Collins & Sorrenson, 2014; ETS, 2017; Esfandiari, 2018; Hill & Liu, 2012; Liu et al., 2009; Phillips, 2015).

ETS and other researchers have conducted copious studies in an attempt to prove the claims that this test represents the contents examinees may face in a real academic setting and that test takers are not required to depend on outside knowledge (Axe et al., 2020; Cohen & Upton, 2006; Esfandiari et al., 2018; Harsch et al., 2007; Hill & Liu, 2012; Liu et al., 2009). Liu et al. (2009) conducted a study on the impact of background knowledge on test takers' scores, utilizing differential item functioning (DIF) and differential bundle functioning (DBF). The study concluded that background knowledge does not have a considerable influence on their scores because all the information needed to understand the reading passages of this test can easily be obtained from the texts. The findings align with ETS's claim that the TOEFL iBT doesn't measure examinees' background knowledge of the topic; instead, it gauges comprehension of the information presented in the given passages (Axe et al., 2020; Collins & Sorrenson, 2014; Phillips, 2015; Putlack et al., 2020). Hill and Liu (2012), in their research, categorized testees into two groups based on their TOEFL iBT scores in an attempt to explore whether or not prior knowledge had an impact on test takers' performance. Those who had high scores were considered high proficiency test takers, whereas those with low scores were regarded as low proficiency test takers. The study, quite different from the result of the abovementioned research, revealed that test takers did not rely largely on prior knowledge when the topics of the passages were general. However, it was, as the study found, difficult to tell whether or not background knowledge could be of major help when the subject was more specific.

In addition, many existing studies in the broader literature have explored the relationship between TOEFL iBT scores and academic success. TOEFL proponents claim that the scores, whether high or low, obtained by test takers can be an indicator of academic performance. Cho and Bridgeman (2012), in their study, found a positive, albeit moderate, correlation between TOEFL iBT scores and *grade point average (henceforth GPA),* which was employed as an indicator of academic achievement. The study also revealed that examinees who obtained higher scores were able to perform better at universities and achieve higher GPA scores; this result led the researchers to conclude that even a moderate

correlation can showcase the fact that TOEFL scores can be a telling indicator of GPA scores. Quite differently, Bridgeman et al. (2015), who in their study divided the testees according to their nationality and the departments they studied in, discovered that TOEFL iBT scores were very positively correlated with GPA. Further, Harsch et al. (2017) found that not only is TOEFL iBT a good indicator of test takers' academic achievement, but it also reveals their ability to utilize and comprehend English in an academic environment. Interestingly, O'Dwyer et al. (2018), dissimilar to the aforementioned studies, discovered that TOEFL iBT can be a telling indicator of students' academic achievement, specifically in language courses, whereas due to the impact of other intervening factors on students' academic performance, this test can only moderately show academic achievement in the courses of the first year at colleges. The studies cited above have all utilized GPA scores as a predictor of academic success to correlate with TOEFL scores. They all indirectly show that a high score on the TOEFL iBT test is a telling indicator of testees' language ability, which is, in turn, interconnected with academic success. This implies that fluency in English provides students with the language skills to perform well in their academic pursuits and the courses they take at university.

However, TOEFL iBT critics claim that this test, albeit designed by professional test designers, is not devoid of pitfalls and may not accurately gauge test takers' ability in English as the topics and the questions are designed in a way that requires test takers to do a multitude of burdening tasks and rely partly, if not mostly, on prior knowledge (Hammad, 2021; Liu, 2011; Ovilia, 2018; Sadighi & Zare, 2006; Sun, 2021; Toker, 2019; Zalha et al., 2020). For instance, Hammad (2021), in her study, tested and interviewed 64 senior students at Al-Aqsa University. The study revealed that the students at this university encountered several problems with the reading section of the TOEFL iBT, namely their lack of background knowledge of the subjects, technical and lengthy passages, limited language ability, etc. The result of the aforementioned study is in line with that of Toker (2019), who maintains that notwithstanding the existence of all the required information in the given passages, examinees still need to have some outside knowledge as these passages possess scores of technical words and sophisticated ideas that require examinees to exert extra efforts to understand and internalize them. Reliance on outside knowledge, as he contends, brings to light the issue of "construct irrelevance variance," which, in turn, undermines and obscures the validity of this test.

In addition, Sun (2021), similar to Toker (2019), asserts that the impact of prior knowledge on reading comprehension is a vital factor to be taken into consideration in the TOEFL iBT reading section. Some test takers, as she argues, resort to utilizing background knowledge to comprehend the passages, which, as a result, calls the reliability and cognitive validity of this test into question because, in such cases, test takers' background knowledge, rather than their reading abilities, determines the result of their scores. The aforementioned claim coincides with a study conducted by Zalha et al. (2020), in which six participants were interviewed concerning their experience with the TOEFL iBT test. The participants responded that while answering the TOEFL iBT reading section, they tended to use their background knowledge to aid them in comprehending the texts.

The impact of background knowledge on listening comprehension of TOEFL iBT is another factor that has been well considered by researchers. For example, Sadighi and Zare (2006) did a study on two groups of students. The first group of students were trained on some of the subjects in the listening section in an attempt to enrich their background knowledge. The other group was not offered any training and lacked background information. In other words, the first group was familiar with background information concerning the topics presented in the listening section, while the second group lacked information regarding the subjects in the listening section. The study discovered that those students who were conversant with the topic and had prior information outperformed those who lacked any information concerning the topics, thereby underscoring the notion that prior information can have an enormous impact on student's ability to understand a text in the listening section. The results of this research are in agreement with those of Ovilia (2018), who conducted a study on ninety-three undergraduate students at a university in Indonesia. The study aimed to investigate whether or not background information can influence TOEFL listening comprehension. It was discovered that students who possessed more outside knowledge about the topics in the listening section were able to obtain higher scores. The results of the aforementioned studies show that outside knowledge can play a vital role in not only reading comprehension but also in listening comprehension.

Moreover, some researchers claim that this test does not predict students' academic achievement, and the scores test takers obtain may not always correlate with their GPAs (Al-Musawi, 2001; Ng, 2007; Vu &Vu, 2013). For example, Al-Musawi (2001) analyzed the TOEFL scores of 90 undergraduate students and their GPAs who studied English as their major and education as a minor at a university in Bahrain in an attempt to investigate their academic performance. The findings of the study revealed that the TOEFL scores did not serve as a good indicator of the student's academic success. The results of this study align with those of Ng (2007), who concentrated on whether or not the TOEFL scores of 433 international students at a college in California predict their academic achievements. The study discovered that the relationship between TOEFL scores and GPA did not prove to be significant. Similarly, Vu and Vu (2013), through a survey, investigated the TOEFL iBT scores of 464 international students and their GPA scores at an American university. The study found no correlation between the students' TOEFL iBT scores and their GPAs. The findings of the studies cited above indicate that obtaining high scores on this test does not confirm that examinees will excel and succeed in their academic performance. This test mainly gauges the language ability of examinees rather than their academic and professional capabilities.

Other factors that have substantial impacts on test takers' performance in the TOEFL iBT test include "motivation, anxiety, ambiguity tolerance, etc."; these traits, which examinees bring with them to the testing centers, are regarded as 'construct-irrelevance variance' (Amiryousefi & Tavakoli, 2011, p. 211; Cheng et al., 2014; Young, 1991). In other words, these traits are regarded as "the potential sources of test bias that can make the obtained scores unrepresentative of the underlying ability that a language test wants to measure and put the whole testing process at stake" (Messick, 1996; Takala & Kaftandjieva, 2000, as cited in Amiryousefi & Tavakoli, 2011, p. 221).

A considerable wealth of literature has been conducted on the issue of whether the score of this test predicts test takers' academic achievements or if this test necessitates examinees to hinge on outside knowledge. However, there has been relatively little literature conducted on investigating if the reading section of this test covers too many technical and infrequent words and the extent to which these technical words; if any, influence test takers' achievements in the exam. In addition to probing into issues pertinent to content and construct validity, this study explores whether or not the reading section of this test is designed to take individual differences into consideration--an issue that has been overlooked by TOEFL critics.

2.2. Variables Affecting Reading Comprehension

Reading, a multifaceted language skill, is a process whereby readers decode and decipher what is written in a text. Until the 1960s, the vast majority of the definitions of reading revolved around this simple description: reading is a process whereby readers read the characters and symbols written in a text to understand them, and it was regarded as a bottom-up process (Alshammari, 2012; Fries, 1945, 1963, 1972; Nguyen, 2012; Rivers, 1968). This definition regards reading as a passive process and maintains that meaning can solely be found in the text, which readers can extract from if they comprehend the whole written text (Zhao & Zhu, 2012). In other words, the aforementioned definition of reading lays considerable emphasis on written texts and disregards the fact that the life experience and background knowledge readers play active roles in interpreting and constructing meaning from written texts.

However, the process involved in reading is much more complicated than most people think. According to Harris and Hodges (1995), while reading, readers construct meaning from a written text through a complex interaction between the readers who carry out the interpretation as well as the message presented in the text. That is, reading is an interactive process whereby readers construct meaning from a text by decoding phonemes, recognizing words, and processing sentences syntactically and semantically, along with their previous experience and background knowledge (Anderson & Pearson, 1984; Carrell et al., 1988; Goodman, 1967; Nguyen, 2012).

Reading, as previously believed to be a process whereby readers only read and understand what is written in a text, can be affected by a throng of factors pertinent to linguistics, "readers' characteristics, nature of materials, and reading tasks" (Alderson, 2009; Zhi-hong, 2007, as cited in Alavi & Bordbar, 2012, p.451). The factors that are related to linguistics include grammatical intricacy, word frequency, and types of topics, whereas the factors pertinent to readers encompass "content schema, metalinguistic knowledge, and metacognitive strategy" (Alderson, 2000, as cited in Kim & Jang, 2009, p. 828). It is vital to evaluate texts and readers together since they are continually interacting, as evaluating them separately will almost surely lead to some degree of distortion (Alderson, 2000).

Grammatical intricacy and word frequency are among those linguistic factors that have tremendous impacts on test takers' comprehension of a text (Li, 2018; Kim & Jang, 2009). A text with more complex structures and low-frequency words causes readers to face difficulty comprehending the text because it requires them to exert extra effort to construct

meaning from it (Bachman, 1990; Syarif, 2018). For example, a study conducted by Cohen et al. (1979) discovered that native Hebrew speakers had difficulty understanding a text written in English because the text contained many complex noun phrases functioning as subjects and objects. This illustrates that text-related and reader-related factors together contribute enormously to comprehension.

Moreover, content schema, which is defined as the cultural knowledge, life experience, and background information students possess regarding a passage, is another factor that has, for decades, sparked a heated debate among researchers and contributes enormously to reading comprehension (Carrell et al., 1988). The background knowledge and the life experience of students, as Anderson et al. (1977) maintain, contribute to the readability of the materials as they aid in decoding and constructing meaning from a text. For example, Alderson and Urquhart (1983, 1984, 1985) conducted a study on whether or not content knowledge and language proficiency aid in understanding a text and found that the type of text determines these factors. They also discovered that background knowledge can have considerable impacts on students to comprehend challenging passages. Whereas texts that are simple and devoid of complicated terminologies and ideas allow students to depend more on language proficiency than background knowledge to understand the text. Sutarsyah's (2009) study is consistent with those of Alderson and Urquhart (1983, 1984, 1985), discovering that incorporating familiar information in a passage leads testees to activate their background knowledge, which, consequently, requires them to exert less efforts to dissect and understand the text. This shows that there is a direct correlation between reading comprehension and the background knowledge readers have about a text.

Metacognitive knowledge, also known as metacognitive awareness, is another variable that immensely contributes to reading comprehension (Kim & Jang, 2009). Metacognitive knowledge refers to "the awareness one has about her or his knowledge and the regulation of learning processes to meet the demands of particular tasks" (Siqueira et al. 2020, p. 2). Readers who use metacognition strategies closely check their reading strategies and engage in critical thought about what is written in a text (Khonarmi & Kojidi, 2011). Readers pass through a multitude of mental processes to both comprehend and construct meaning from a text because reading is a process of decoding and deciphering what is in a text, which depends largely on metacognition (Grabe, 2009; Khonarmi & Kojidi, 2011; Miller, 2017).

The aforesaid studies show that reading relies largely on a range of factors related to both texts and readers. In other words, the factors and/or variables that contribute hugely to comprehension include word frequency, grammatical intricacy, content schema, and metacognitive knowledge. Because readers and texts depend on one another continually, evaluating them separately almost certainly leads to some degree of distortion (Alderson, 2000; 1985; Kim & Jang, 2009; Sutarsyah, 2009).

2.3. Construct-irrelevant Variance and Construct Under-representation

Messick (1989) identifies two threats to the validity of a test: construct-irrelevant variance and construct under-representation. Construct-irrelevant variance occurs when the constructs of a test are broad and irrelevant factors influence the intended construct of a test. For example, if a biology exam aimed at measuring students' ability in biology knowledge is

influenced by test takers' reading comprehension skills, this is regarded as construct-irrelevant variance. Further, such factors as test anxiety, cultural bias, test environment, and extraneous materials pose a threat to the true measurement of the intended construct, thereby reducing the validity of the test. Messick (1995) then subdivides construct-irrelevant variance into construct-irrelevant difficulty and construct-irrelevant ease. The former occurs when test items are designed in a way that makes the test more difficult than it should be. For example, a passage with too many technical words and convoluted sentences contaminates the intended construct of the test, which, consequently, lends invalidity and unfairness to the test. On the contrary, the latter occurs when test items are designed in a way that makes the test easier and less challenging than it should be. For instance, if the items of a question are so simple and designed in a way that cues savvy students to the correct answers, this is regarded as construct-irrelevant easiness.

In contrast, construct under-representation occurs when the construct of a test is narrow and does not cover the entire domain of the intended construct. For example, if a reading test is designed in a way that gauges only vocabulary knowledge but ignores other aspects related to reading comprehension skills, this test is contaminated as it fails to reflect the intended construct of reading comprehension tests (Messick, 1989).

3. Methodology

The present study utilizes a combination of quantitative and qualitative research methods to analyze the selected data. It examines the representation of various academic topics in the TOEFL iBT reading section and investigates how effectively this section measures test-takers' ability to understand and utilize academic English. Further, the study aims to both investigate whether or not the TOEFL iBT reading section includes too many technical and infrequent vocabulary items and the extent to which technical and low-frequency words, if any, in the TOELE iBT reading passages (including in the vocabulary questions) have a negative impact on test takers' performance in the test. Further, it also attempts to explore the extent to which these individual differences may or may not influence students' achievement in the test

3.1. Procedures and Materials

A total of 209 authentic passages were selected from five different books and categorized into four distinct categories: hard sciences, social sciences, arts and humanities, and business and management. The selected books include:

- The Complete Guide to the TOEFL Test: iBT Edition (2007) by Rogers
- McGraw-Hill Education TOEFL iBT (2014) by Collins and Sorrenson
- Longman Preparation Course for the TOEFL iBT Test (2015) by Phillips
- The Official Guide to the TOEFL® Test (2017) by ETS
- Decoding TOEFL iBT: Actual Test, Reading 2 (2020) by Putlack et al.

The distribution of reading passages from each book is as follows:

Book Title	Year	Number of Passages
The Complete Guide to the TOEFL Test: iBT Edition	2007	75
McGraw-Hill Education TOEFL iBT	2014	17
Longman Preparation Course for the TOEFL iBT Test	2015	72
The Official Guide to the TOEFL® Test	2017	18
Decoding TOEFL iBT: Actual Test, Reading 2	2020	27

The selected books were published in different years to ensure research validity and add variety to the reading test samples.

3.2. Categorization of Passages

The classification of the reading passages was conducted according to the categorizations proposed by Collins and Sorrenson (2014):

- **Hard Sciences**: Life sciences, physical sciences, earth and space sciences, computer science, and engineering.
- **Social Sciences**: Geography, sociology, anthropology, psychology, and economics.
- **Arts and Humanities**: Literature, art and art history, music, and music history.
- **Business and Management**: Case studies.

The study uses Excel as a data analysis tool to organize the data and calculate the percentage distribution of the passages into each category.

3.2.1. Expert Involvement

After their classifications, the 209 reading passages were handed to two IELTS and TOEFL trainers, who are also university instructors, to ensure precise classification and substantiate the validity of the categorization. Their involvement in ensuring the precise categorization added another layer of the validity and reliability to the classification process.

3.3. Word Frequency Analysis

To examine whether or not the TOEFL iBT reading section covers too many technical and infrequent vocabulary items, the word frequency across all four categories in all the selected passages was investigated based on the Collins Dictionary (2023), which derives its data from the *Collins Corpus* and *the Bank of English.* This dictionary classifies word frequency into five types:

- **Extremely Common**: Among the 1,000 most widely used words.
- **Very Common**: Among the 4,000 commonly used words.
- **In Common Usage**: Among the 10,000 widely used words.
- **Used Occasionally**: Among the 30,000 commonly used words.
- **Used Rarely**: Below the 50% threshold of widely employed words.

The study focuses on identifying words categorized as "Used Occasionally" and "Used Rarely," as these are considered low-frequency words. The TOEFL iBT reading section typically includes three or four passages, each approximately 700 words long (Phillips, 2015). This study examines the passages specifically to identify those words that fall into the categories of *Used Occasionally* and *Used Rarely*, as these two categories are regarded as low-frequency words. Excel as a data analysis tool is employed to organize the data and calculate the percentage distribution of the word frequency into each category put forward by Collins dictionary.

3.4. Vocabulary Question Analysis

In addition, to identify word frequency in vocabulary questions, the vocabularies chosen to test examinees in the MCQs in these five books encompassed 287 words categorized into five types classified by the Collins Dictionary (2023) mentioned above. The reason behind categorizing the words based on their frequency of use in vocabulary question types was to investigate whether or not the vocabularies utilized to test examinees in the vocabulary questions are high- or low-frequency words in the reading context. Excel as a data analysis tool is employed to organize the data and calculate the percentage distribution of the word frequency into each category put forward by Collins dictionary.

3.5. Individual Differences and Multiple-Choice Questions

The reading passages were evaluated qualitatively based on the quantitative data results using Messick's (1989) concepts of construct under-representation and construct irrelevant variance threats. Also, the quantitative was analyzed to determine whether the TOEFL iBT has accounted for the topic effect, knowledge schemata, when designing and distributing the passages in the reading section. Moreover, through the prism of the Messick's two concepts and knowledge schemata, the reading passages were evaluated qualitatively based on the quantitative data results to determine whether or not they took individual differences into account when assessing examinees' ability to understand a text.

Multiple-choice questions from the selected reading passages were examined to see if the stems and list of options posed a threat to the test's validity or accurately assessed students' reading comprehension skills. Haladyna's (2022) criteria and guidelines for creating and designing MSCQs were used to identify such technical flaws as convoluted sentences, grammatical cues and mistakes, long answers, horizontality, word repeats, negation in the stem and options, ridiculous choices, and absolute terms, which, if present in the stems and list of options, pose a serious threat to the validity of the questions.

4. Results and Discussions

4.1. Validity and Topic Effect

This study investigated whether or not the reading section of the TOEFL iBT is representative of all the types of reading topics examinees face in real academic life. For this reason, 209 reading passages were selected from five different books and classified into four categories: hard sciences, social sciences, arts and humanities, and business and management. The paper also investigated the extent to which the reading section of this test measures test

takers' ability to utilize and understand academic English topics. The distribution of the 209 passages chosen for this study is clearly shown in the following bar chart:

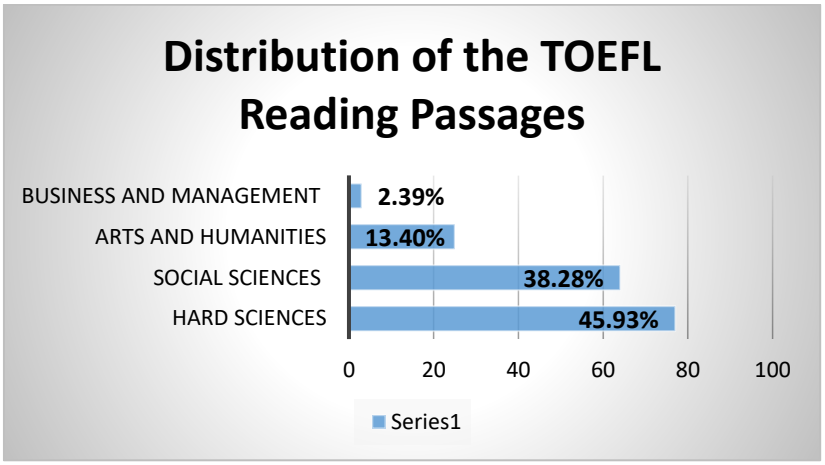


Figure 1: Distribution of the TOEFL Reading Passages

As shown above, overall, the passages that fall into the categories of hard sciences and social sciences together account for over half of the reading section passages. The hard science subjects constitute 45.93%, whereas the social science passages amount to 38.28%; in total, they make up four-fifths of the reading passages combined together. However, arts and humanities and business and management account for 13.40% and 2.39%, respectively, representing less than 20% of all passages in the reading part. Upon closer inspection, the chart indicates that the highest percentage recorded is for the hard science passages, whereas the business and management passages constitute the lowest percentage with over two per cent.

A quick look at the reading section shows that this section does not equally cover all the academic areas, such as history, art, music, business, and so forth. In other words, the domain of the hard sciences is much wider than the domain of those topics related to the social sciences, arts and humanities, and business and management. This result is close to the work of Li (2018), who discovered that 41.2% of the TOEFL iBT passages are related to the hard sciences, whereas 58.2% fall into the categories of social sciences, arts and humanities, and business and management. Putlack et al. (2020), in line with the result of the current study, maintain that the 2019 TOEFL iBT reading section went through several modifications, and after these modifications, this section now places considerable emphasis on those passages related to *hard sciences*, namely biology, zoology, etc. This result is a telling indicator of how the reading section favors some test takers over others and how this test seemingly presupposes that all students who can read English and who are somehow determined to be of equal proficiency should do equally well on the test regardless of whether they have a background on the topics or not.

ETS maintains that the content domain of the TOEFL iBT reading section reflects a blend of different topics, such as business, hard and social sciences, and the arts (Axe et al., 2020; ETS, 2017). The reason behind including a mixture of different topics, as maintained by ETS, is to uphold the content validity of the reading section in order for the test to equally tap into all facets of a theoretical concept (Collins and Sorrenson, 2014, p. 106). However,

this section, as illustrated in the bar chart, favors hard science topics over those about the arts or business. Favoring some topics over others brings up the issues of topic effect and construct under-representation, which usually pose a serious threat to the validity of reading tests. This goes against the definition of content validity, which says that a test possesses "content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned" (Hughes, 2003, p. 26).

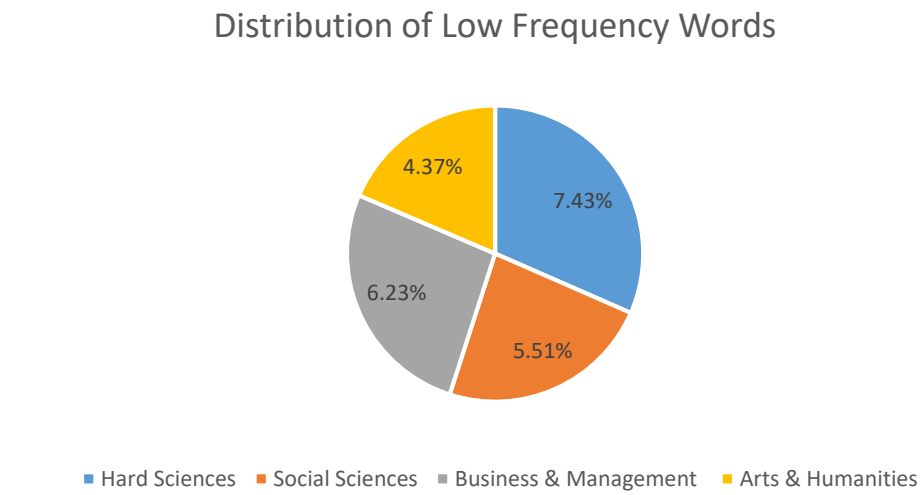
The hard science topics, as illustrated in the chart, show that the content and construct validity of this test are very high, especially for students who have studied such topics in their schools. On one hand, it could well be argued that the reading section does reflect some future tasks that students will or may perform in their future academic studies, as test designers presuppose that students have some knowledge of these tasks in order to answer them correctly. On the other hand, science-related subjects may be completely foreign and thus unfair to populations of international students whose curricula either ignore these topics or even deny their validity in their own dominant cosmogony. This brings up the issue of construct irrelevance variance as students with background knowledge of hard science subjects outperform those less familiar with science-related topics. In other words, showing preference for certain topics over others can result in the issue of "construct irrelevance variance", causing some test takers to excel in reading tests while others struggle and underperform in exams. That is, some students would probably do much better on some topics than others, depending on their own personal background with the subject or lack of background with the subject. For example, a person who studies biology performs better on those topics related to hard sciences than a person who speaks English fluently and studies music or literature (Anderson & Lynch, 2000; Gebhard, 2000; Othman & Vanathas, 2017).

The above-discussed view is well supported by Alderson and Ulquhart (1983, 1984, 1985), Clapham (1998), Dechant (1991), and Toker (2019), who contend that readers understand the meaning of a text with the help of the knowledge they have concerning the topic. Lee (2011) also reinforces that prior knowledge plays an important role in making readers actively engaged in deciphering and decoding a text. Schmidt-Rinehart (1994) emphasizes the importance of topic familiarity for students' performance and achievement in a test. In his study, the researcher provided a group of students with two different passages. The students had prior knowledge about one passage, whereas the second passage was new to them. The result showed that the students had better recall of the familiar passage, which consequently led them to score significantly higher on the familiar topic. This result allowed the author to conclude that not only does the topic effect have a considerable effect on reading comprehension, but it also has an enormous influence on listening comprehension. The result of this study coincides with the studies conducted by Chiang and Dunkel (1992), Ovilia (2018), Priebe et al. (2012), and Sadighi and Zare (2006), which discovered that students scored significantly higher on topics familiar to them.

4.2. Word Frequency in the TOEFL iBT Reading Passages

It is argued that one of the issues that poses a threat to the construct validity of this test is that the passages in the reading section are so convolutedly written and/or discussed that it requires test takers to possess some prior knowledge to comprehend the passages

believed to be imbued with technical and low-frequency words (Asrida & Fitrawati, 2019; Maizarah, 2019; Zarnis, 2020). However, ETS staunchly rebuts this view, claiming that the TOEFL reading section, which usually covers familiar topics, does not encompass many low-frequency words. In their regard, they claim that test takers are not required to have prior knowledge since all the information can easily be obtained from the texts (Educational Testing Service, 2023). The TOELE iBT reading section consists of three or four passages, and each passage is around 700 words (Phillips, 2015). The percentage of low-frequency words distributed amongst the four types of reading passages is as follows:



:

Figure 2: Distribution of Low-frequency Words

As it can be seen, 7.43% of the words in those passages related to the hard sciences fall into the categories of Used Occasionally and Used Rarely, whereas infrequent words in those passages relevant to the arts and humanities account for only 4.37%. Following the hard science passages are the business and management passages and the social science passages, which comprise 6.23% and 5.51%, respectively. This pie chart shows that the hard science passages constitute the highest number of low-frequency words as opposed to the art and humanities passages, which consist of the least number of infrequent words

This result supports ETS's claim that TOEFL iBT reading passages do not cover many low-frequency words, and a glossary is available to provide examinees with the meaning of those words that are technical or utilized rarely (Educational Testing Service, 2023). In other words, the result shows that each passage contains less than 10 low-frequency words out of 700 words, and this relatively small number does not have severe impacts on test takers' comprehension of the given passages. ETS's major attempt is to refrain from designing a reading comprehension test that is teeming with technical and infrequent words in an effort to substantiate the construct validity of the test (Educational Testing Service, 2024). For example:

A) It is obvious that cetaceans—whales, porpoises, and dolphins—are mammals. They breathe through lungs, not through gills, and give birth to the living young. Their streamlined bodies, the absence of hind legs and the presence of fluke[1] and blowwhole[2] cannot disguise their affinities with land dwelling mammals. However, unlike the cases

of sea otters and pinnipeds (seals, sea lions, and walruses, whose limbs are functional both on land and at sea), it is not easy to envision what the first whales looked like. Extinct but already fully marine cetaceans are known from the fossil record. (ETS, 2017, p. 60)

B) Parasitic plants are plants that survive by using food from a host plant rather than producing their own food from the sun's energy. Because they do not need sunlight to survive, parasitic plants are generally found in shaded areas as opposed to areas exposed to direct sunlight. The plants can be classified in various ways; one of the most prevalent methods is by determining whether the plant depends wholly on its host (holoparasite) or has some degree of photosynthesis[1] ability (hemiparasite), which allows it to provide some of the own nutrients when necessary. (Phillips, 2015, p.19)

The two aforementioned passages talk about the origins of cetaceans and parasitic plants, respectively. A closer examination of the two excerpts reveals that both passages contain a small number of infrequent words, such as "cetaceans," "gill," "hind," "fluke," "blowhole," "pinnipeds," and "photosynthesis." The words "cetacean," "blowhole," and "pinniped", according to the Collins Dictionary (2023), are regarded as "Used Rarely," because they are placed among the lower 50% of commonly used words, whereas "gill," "hind," and "fluke" are classified as "Used Occasionally," meaning they fall within the 30,000 most commonly used words.

The abovementioned excerpts show that TOEFL iBT designers are likely to refrain from employing exhaustive technical terms in their tests. Rather, they opt to provide clarifications for such technical or infrequent words either included within brackets in the passage or they may have them explained in footnotes. For example, the words "fluke," "blowhole," and "photosynthesis" are explained in footnotes, while "pinniped" and "cetacean" are clarified within brackets in the passages (Educational Testing Service, 2024, Para. 3; Phillips, 2015, p.19; Rogers, 2007, p. 3)

TOEFL iBT designers embrace such an approach in an endeavor to place considerable importance on not only test takers' ability to identify the meanings of the words but also on their capacity to understand the concepts and ideas mentioned within the passages provided. The designers of the selected passages have chosen those passages that contain neither too simple nor challenging words, thereby eschewing construct-irrelevant difficulty and construct-irrelevant easiness that are subsumed under construct-irrelevance variance. Test designers aim to design an unbiased form of reading assessment that accurately and precisely measures test takers' comprehension of the subject matter (Educational Testing Service, 2023, para 3).

This aforementioned view is supported by Bachman (1990), who maintains that test takers can better comprehend and internalize those reading passages with high-frequency words as opposed to passages with more low-frequency words. Bachman's claim is in alignment with a multitude of psycholinguistic theories. Morton (1979), who developed the Logogen Model, maintains that high-frequency words, unlike low-frequency words, put less strain on the brain to recognize, decode, and retrieve from the mental lexicon. This is because high-frequency words, unlike low-frequency words, have a higher threshold. According to Forster (1976), who developed the Autonomous Search Model, words are systematically

arranged in the lexicon in which the low-frequency words are located at the back, whereas the high-frequency words are located in the front, meaning a text with more high-frequency words requires less power on the brain and less information to comprehend and internalize.

Further, low-frequency words, according to Gough (1984), need to go through two different processes in the brain, namely the phonological process and the comprehension process. These two processes require test takers to exert more effort to understand a passage. Whereas high-frequency words do not need to go through the phonological process in the brain; they directly go to the comprehension process, which, in turn, causes test takers to exert less effort to understand a passage that contains more high-frequency words (Gough, 1984).

4.3. Word Frequency in Vocabulary Questions

Vocabulary questions in the TOEFL iBT reading section measure test takers' comprehension of English vocabulary or phrases in a given context. In other words, the purpose behind this type of question is to enable test takers to utilize context as a clue to understand those words that have more than one meaning (Collins & Sorrenson, 2014; Phillips, 2015; Putlack et al., 2020). The words in the questions selected for this study encompass 287 words that have been classified into five types based on the categories of the Collins Dictionary (2023): Extremely Common, Very Common, In Common Usage, Used Occasionally, and Used Rarely. The rationale behind categorizing the words based on their frequency of use is to investigate whether or not the vocabulary utilized to test examinees in the vocabulary questions are high- or low-frequency words. The following pie chart shows the percentage of the frequency of the vocabulary tested in the reading section:

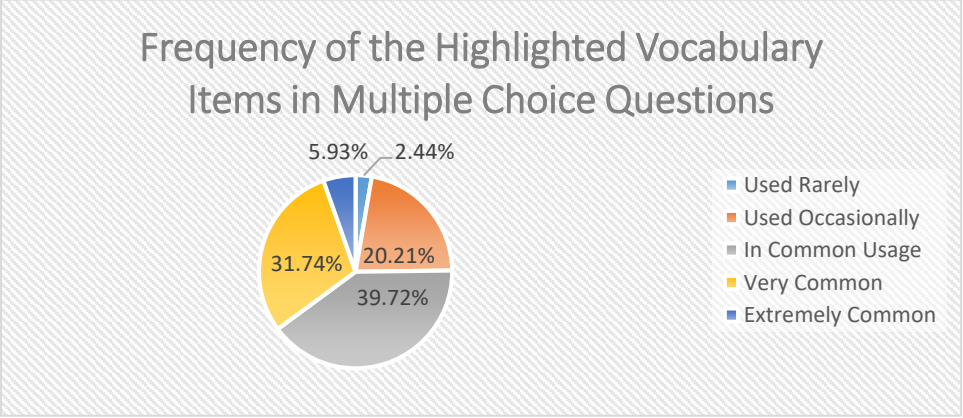


Figure 3: Frequency of Highlighted Vocabulary Items in MCQs

As the pie chart illustrates, the words that fall into the category of In Common Usage account for 39.72%, whereas the Used Rarely (2.44%) and the Extremely Common (5.93%) words together make up below 10% of all the vocabulary tested. Following the In Common Usage vocabulary items are the Very Common and the Used Occasionally words, which constitute 31.74% and 20.21%, respectively. Upon closer examination, it is clear that those words that belong to the category of In Common Usage represent the largest portion of the highlighted words in the reading section, whereas those words that come under the categories of Used Rarely and Extremely Common comprise the least number of highlighted words.

The result shows that most of the vocabulary items used to test examinees are in Common Usage which is regarded as words of high frequency. In other words, those words that belong to the categories of In Common Usage, Very Common, and Extremely Common together account for 77.39% of all the highlighted words, meaning most of the words chosen to test examinees are high-frequency words. The pie chart indicates that ETS opts to refrain from incorporating low-frequency words into the vocabulary questions in an attempt to substantiate the validity of this test and eschew construct-irrelevant difficulty threat. Rather, it utilizes those common words that possess multiple meanings to allow testees to select the correct option based on the context of the words in the passages. Thus, TOEFL iBT adopted such approach of selecting those common words that have multiple meanings--in an effort to avoid construct-irrelevance easiness (Collins & Sorrenson, 2014; Phillips, 2015; Putlack et al., 2020).

4.4. The Validity of Multiple-Choice Questions in TOEFL iBT

The TOEFL iBT reading section comprises three or four passages, and each passage consists of 12–14 questions (Collins & Sorrenson, 2014). All the questions are multiple-choice questions, which are a kind of question or a form of assessment in which test takers are provided with a problem, known as a stem, and a list of options (Al-Faris et al., 2010; Torres et al., 2011). Test designers should consider this fact that MCQs pose a potential threat to the validity of a test if they fail to both reflect students' ability in reading comprehension and enable them to construct meaning from the text (Al-Faris et al., 2010; Brookhart, 2015; Haladyna, 2022). In other words, MSCs should not be designed in a way that allows examinees to depend entirely on test strategies to get good grades (Al-Faris et al., 2010; Baghaei & Amrahi, 2011; Brookhart, 2015; Rodriguez, 2005). Rather, the questions, as Toker (2019) maintains, should allow students to depend on generating meaning from the texts using meaning-making tools rather than relying mostly on particular test methods or test-wiseness to answer the questions.

Given the questions in the passages reviewed for this study, the structures of the questions are well-constructed and clearly and concisely stated. The questions and the list of options are not verbose because, according to Haladyna (2022), "the linguistic complexity of an item stem might challenge the student unfairly" (p.14). In other words, Brame (2013) maintains that "items that are excessively wordy assess students' reading ability rather than their attainment of the learning objective" (para. 10). That is, if the questions and the list of options are long-winded, the objective of the questions shifts from gauging test takers' understanding of the reading passage to reading speed, which may, consequently, lead them to exert considerable effort to understand the questions. Further, the stems and distractors in all the passages selected for this study are devoid of convoluted sentence structures as well as highly specialized and technical words, as convoluted structures and infrequent words prevent examinees from easily comprehending the questions, which may, consequently, lead the stem and the distractors to test examinees' vocabulary knowledge and obscure the true purpose of the stems and distractors, which recognize the question or problem and the answer, respectively (Haladyna, 2022).

In addition, the questions in the passages reviewed for this research are devoid of such technical flaws (mentioned by Haladyna, 2022) as grammatical cues, long answers, and word repetition because test designers have designed the questions in a way that doesn't provide test takers with any cues to choose the correct options unless test takers carefully read and understand the passages. Moreover, the vast majority of the stems in the selected passages are devoid of such absolute terms or negative verbiage as 'never, always, none, rarely, etc.' as these terms cue savvy students to the correct options, which, in turn, obscure and undermine the validity of the questions (Begum, 2012; Haladyna, 2022; Torres et al., 2011). Further, the stems are all stated either in incomplete statements, direct questions, or in a positive sentence. All the stems are followed by one correct answer and three distractors, which ensure faster reading and response times for examinees. For example:

- **Which of the following can be inferred about peppers sold in supermarkets?**

A. Spicier peppers are more expensive than less spicy varieties.
B. Peppers with Scoville scores higher than 200,000 units are not safe to eat.
C. Peppers with Scoville scores higher than 200,000 are available less frequently than habanero peppers.
D. Peppers hotter than habanero peppers are never sold in supermarkets.

(Collins & Sorrenson, 2014, p. 19)

This excerpt is a question from a passage that talks about human perceptions of food flavor (Collins & Sorrenson, 2014). A closer examination of the excerpt reveals that the stem is not a "garden path sentence" (Bever, 1970), which, if present in the questions, jeopardizes the validity of the test and forces test takers to read and reinterpret them numerous times. Rather, the stem is clearly stated as a direct question and devoid of any grammatical cues, convoluted structures, absolute terms, or negative verbiages. Moreover, distractors A, B, and D are devoid of convoluted structures and do not contain any ridiculous options or grammatical cues to cue savvy testees to the correct option. The same holds true for the correct option, which is C. Further, the correct option neither provides any words or phrases that resemble the stem, nor does it contain more specific and detailed information than the distractors. The options are homogenous in content in that they all talk about peppers because heterogeneous options can easily cue test wise students to the correct option. Further, the stem and the list of options are all designed horizontally and do not contain those infrequent terms that deter test takers from understanding the question. These are all in alignment with the guidelines and critera set by Haladyna (2022).

The MCQs in the reading section have strong validity as they are designed in a way that measures testees' abilities in understanding a passage, such as vocabulary, references, sentence restatements, sentence insertion, inferences, fact and negative fact, and overall organization of ideas (Phillips, 2015). The questions do not focus on measuring one specific area of test takers' ability to understand a passage. Rather, they act like an umbrella under which all the forms of measurement and assessment are well and concisely presented.

4.5. Individual Differences

Another issue that calls the construct validity of this test into question is individual differences. Such differences encompass scores of factors, namely language proficiency,

culture, gender, learning style, personalities, and social behavior, which distinguish one individual from another (Ahmadjavaheri & Zeraatpishe, 2020; Bachman & Palmer, 1996). Therefore, the purpose of education, curriculum, and testing should be to take these aforesaid individual differences into account. Moreover, test design should carefully consider students' individual differences and assess their performance accurately, including their emotional, behavioral, and social skills. In other words, test administrators should refrain from designing tests that merely assess students' test-taking abilities (Ahmadjavaheri & Zeraatpishe, 2020).

However, many TOEFL critics claim that individual differences and such things are not officially major considerations made by TOEFL administrators (Amiryousefi & Tavakoli, 2011; Toker, 2019). As mentioned above, the fact that 45.93% of the passages center on scientific subjects presupposes that all students can *equally* do well on these subjects regardless of whether they studied them before or not. In other words, the test designers who have chosen the passages reviewed for this study seem to have paid little attention to the reality that designing a test without taking individual differences into consideration may divert the outcomes of the test in other directions, which, as a result, causes the test not to be a good reflection of students' ability and their test scores. The emphasis that the reading section places on science-related subjects leads to an "construct under-representation threat" (Messick, 1989), which, in turn, causes certain examinees to excel in reading performance and receive higher grades.

To conduct a comprehensive and unbiased evaluation of students' academic proficiency through testing, it is critical for test administrators to recognize that students' backgrounds differ and that this difference influences the outcome of their test results (Ahmadjavaheri & Zeraatpishe, 2020; Kubat, 2018). However, the reading section seems to presuppose some standardized paradigm from which all students can equally depart. To put it differently, the content of this section presupposes that all students have an equal departure point with the information or content of the question, and that if this were true, then one would assume the content and construct validity to be very high indeed. But one must ask, "Is this true?" The answer is "no," as the reading section gives some test takers with more knowledge of a topic a hand-up over fellow test takers who do not know the subject. This indirectly indicates that the reading section fails to consider the fact that students are different in terms of learning styles, personalities, gender, cultural background, prior knowledge, and physical characteristics. By ignoring these essential differences, which, in turn, leads to an "under-representation threat" (Messick, 1989), this exam yields inadvertent consequences that may badly influence students' test achievement. The claims mentioned above are in alignment with those of Amiryousefi and Tavakoli (2011), who in their study found that the TOEFL iBT writing section favors those testees who possess musical intelligence, whereas the listening section is biased towards those examinees who have kinesthetic intelligence.

To provide a comprehensive and fair evaluation of students' academic proficiency through testing, it is important for test administrators to take into account such influential factors as students' gender, cognitive aptitude, perceptual abilities, physical characteristics, and preferred learning styles (Ahmadjavaheri & Zeraatpishe, 2020; Kubat, 2018). If we add to this reality that we all have different intelligences, as Dörnyei and Skehan (2003) and Dörnyei (2009) argue, and the time and stress factors that Krashen (1982) has shown affect

learning outcomes, proficiencies, and test scores of virtually everyone, then the question of the efficacy of the appropriateness and relevancy of the construct validity of these types of passages being applied to all students *universally* and *equally* becomes problematic. For example, one issue that brings the construct of this test into question is that it focuses on MCQs, despite their clear and well-stated structures and forms. The test's excessive focus on MCQs, which brings up the issue of construct under-representation and underrepresents other significant aspects of academic proficiency, may not correspond to all students' various learning preferences and talents. In other words, overreliance on MCQs, which has a negative impact on the test's construct, undermines the ability of those test takers whose strengths lie in essays or open-ended questions. Grandt (1987), in his study, supports the aforesaid view, asserting that test designers should take individual differences into account. He found that female students generally outperformed male students on open-ended answers, whereas male students did perform better on MCQs. The findings of the aforesaid study are consistent with those of Ramos and Lambating (1996), and Pekkarinen (2015), who discovered that female students tended to skip more MCQs in comparison with male students. This shows that overreliance on MCQs in this test, despite being precisely created by experienced and professional test designers, obscures test takers' true aptitude, as prior studies indicated that some students do not applaud MCQs. This also raises the issue of "construct irrelevant variance" (Messick, 1989; Toker, 2019), which overshadows test takers' inner ability and allows the external factor discovered by the aforementioned studies and the ones mentioned at the beginning of the paragraph to divert the test results in other directions.

Moreover, certain test takers with effective fluency may have problems answering the questions related to the reading section of this test. This is applicable to slower readers or learners, as they might struggle to complete the reading section on time owing to both the length and technicality of the passages. To put it another way, TOEFL iBT may assist those test takers with dyslexia by demanding medical evidence. However, what about those slower learners who are intelligent but take longer to comprehend and evaluate passages? This issue, which contaminates the test’s constructs and leads to individual differences as well as a construct irrelevant variance threat, has been overlooked by TOEFL iBT designers. Daneman and Carpenter (1980), in line with the aforesaid claim, found that good readers undertake computations when reading in a faster and more effective manner due to their high working memory level. Poor readers, nonetheless, find it challenging to perform such computations when they read owing to their relatively lower working memory capacity. This qualitative attribute, i.e., working memory, sets both good and poor readers apart. The findings of this article suggest that underachieving readers may be involved in basically different reading patterns that are less effective than those of good readers who perform well-practiced and better reading strategies (Daneman & Carpenter, 1980).

In the TOEFL iBT test, instead of focusing on recollection of the knowledge acquired in traditional learning environments, considerable importance is placed on such critical thinking skills as reasoning, inference-making, and sentence restatement (Phillips, 2015). Therefore, overreliance on critical thinking can result in great difficulty for slower learners who seek to do the reading section within the allotted time. This can lead to individual

differences in performance, which, in turn, conceal the true reading abilities of examinees and ultimately call into question the accuracy and validity of the test.

Conclusion

This study examined the TOEFL iBT reading section to determine if it accurately represents the entire domain test takers encounter in an academic environment and whether the TOEFL iBT reading section passages take individual differences into account and accurately measure test takers' comprehension abilities.

The results of this research indicated that the passages in the reading section mostly revolve around hard science topics, with little attention given to topics about business and management. As a result, this may favor those test takers who specialize in or are familiar with topics related to hard sciences over those who have not studied these subjects in school. This, consequently, overshadows the true ability and proficiency of test takers, including those who are considered slower learners. The issue of how the passages are unequally distributed should be taken into serious consideration in order to accurately assess different test takers from a range of different academic fields.

In addition, the excessive reliance on multiple-choice questions contaminates the test construct as it doesn't account for testees' learning styles and preferences. Thus, different types of topics and questions need to be incorporated into this test to accurately assess and evaluate examinees with varying abilities.

However, most of the words in the reading passages and vocabulary questions are regarded as high-frequency words, which, in turn, contribute to the readability of the passages. This ensures that low-frequency words do not hamper test takers, which aligns with the main objective of reading test goal: to gauge students' ability to understand the entire passage and eliminate potential comprehension issues caused by unfamiliar vocabulary.

The findings also revealed that the multiple-choice questions are appropriately designed and devoid of grammatical errors, word repetition, ambiguous words, absolute terms, grammatical cues, and convoluted sentence structures, which substantiates the validity and robustness of multiple-choice questions. This is well aligned with the goal of multiple-choice questions in reading tests, which is to gauge and test students' comprehension of entire passages.

While the study underscores the idea that designing valid MCQs and eschewing passages teeming with low-frequency words are necessary, it also recommends the inclusion of a variety of different topics and questions in the reading section, which better reflects students' different backgrounds learning styles. Further, time constraints, which prove to be obstacles for those who are deemed slower learners, need to be modified to accommodate individual differences among examinees.

This study significantly contributes to scholarship concerning the TOEFL iBT reading section but exhibits limitations, including the limited number of passages and books examined. Future research could expand by including a wider variety of reading passages

from different books published over an extended period, offering a more comprehensive approach for more accurate results. Future trends could also account for cultural and pedagogical diversity, impacting students' test results. Additionally, highlighting the experiences of slower learners during the TOEFL iBT test--specifically how time constraints affect their scores compared to others without such limitations-- could provide valuable insights.

References

- Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct irrelevant factors on the validity of reading comprehension test. *International Journal of Language Testing, 10*(1), 1-10.
- Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct-irrelevant factors on the validity of reading comprehension tests. *International Journal of Language Testing, 10*(1), 1-10.
- Alavi, S. M., & Bordbar, S. (2012). A Closer look at reading strategy use in reading section of TOEFL iBT. *Theory and Practice in Language Studies (2)*3.
- Alderson, J.C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing, 26*(4), 621–631.
- Alderson, J. C., & Urquhart, A. H. (1983). The effect of student background discipline on comprehension: A pilot study. In A. Hughes & D. Porter (Ed.), *Current developments in language testing* (pp. 121–127). London: Academic Press.
- Alderson, J. C., & Urquhart, A. H. (1984). Student discipline and reading comprehension. In T. Culhane (Ed.), *Practice and problems in language testing*. University of Essex Occasional Papers
- Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192–204.
- Al-Faris, E. A., Alorainy, I. A., Abdel-Hameed, A. A., & Al-Rukban, M. O. (2010). A practical discussion to avoid common pitfalls when constructing multiple choice questions items. *J Family Community Med, 17*(2), 96-102.
- Al-Musawi, N. (2001). The validity of scores on TOEFL and FCE for predicting students' success at the university. *Dirasat: Educational Science, 28*(1).
- Alshammari, H. A. M. (2012). Effects of time constraint on second language reading comprehension [MA Thesis, Southern Illinois University/Carbondale].
- Amiryousefi, M., & Tavakoli, M. (2011). The relationship between test anxiety, motivation and MI and the TOEFL iBT reading, listening and writing scores. *Procedia - Social and Behavioral Sciences, 15*, 210-214.
- Anderson, A., & Lynch, T. (2000). *Listening.* Oxford: Oxford University Press.
- Anderson, R.C., & Pearson, P.D (1984). A schematic-theoretic view of basic processes in reading. In P.D. Pearson, M. Kamil, R.Barr, & P.Mosenthal (Ed.), *Handbook of reading research* (pp.255-291). New York: Longman.
- Anderson, R. C., Pichert, J. W., & Shirey, L. L. (1979). *Effects of the reader's schema at different point in time*. Urbana-Champaign, IL: University of Illinois, Center for the Study of Reading.
- Axe, et al. (2020). Validity evidence supporting the interpretation and use of TOEFL iBT scores. (TOEFL iBT Research Report Vo. 4). Princeton, NJ: Educational Testing Service.
- Asrida, R., & Fitrawati, F. (2019). The Difficulties of English department students

at Universitas Negeri Padang in answering reading section of TOEFL.
Journal of English Language Teaching, 8(4), 496-503–503.

- Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.
- Bachman, L., Davidson, F., Ryan, K., & Chol, Inn. (1995). *An investigation into the comparability of two tests of English as a foreign language*. Cambridge, Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford: Oxford University Press.

- Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple-choice items. *Psychological Test and Assessment Modeling, 53*(2), 192-211.
- Begum, T. (2012). A guideline on developing effective multiple-choice questions and construction of single best answer format. *Journal of Bangladesh College of Physicians and Surgeons, 30*(3), 159-166.
- Bever, T.G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language*. (pp. 279-362). New York: Wiley.
- Brame, C. (2013). *Writing good multiple choice test questions*. Center for Teaching Vanderbilt University.
- Bridgeman, B., Cho, Y., & DiPietro, S. (2015). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing, 33*(3), 307–318.
- Brookhart, S. M. (2015). Making the most of multiple choice. *Educational Leadership, 73*(1), 36-39.
- Carrell, P., Devine, J., & Eskey, D. (1988). *Interactive approaches to second language reading*. Cambridge: Cambridge University Press.
- Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and test anxiety in test performance across three testing contexts: The CAEL, CET, and GEPT. *TESOL Quarterly, 48*(2), 300–330.
- Chiang, C.C., & Dunkel, P. (1992). The Effect of speech modification, prior knowledge and listening proficiency on EFL lecture learning. *TESOL Quartely. 26* (2), 345- 374.
- Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421–442.
- Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 141–168). Mahwah, NJ: Lawrence Erlbaum
- Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1979). Reading English for specialized purposes: Discourse analysis and the use of student informants. *TESOL Quarterly, 13*, 551–564.
- Cohen, A. D. & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. TOEFL Monograph Series, MS – 33. Princeton, NJ: Educational Testing Service.
- Collins and Sorrenson. (2014). *McGraw-Hill education TOEFL iBT*. New York: McGraw-Hill Education.
- Collins English Dictionary. (2023). Available from: https://www.collinsdictionary.com/ [Accessed 2 May 2023]
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6
- Dechant, E. (1991). *Understanding and teaching reading: An interactive model*. Hillsdale, NJ: Lawrence Erlbaum.
- Dörnyei, Z. (2009). *The Psychology of second language acquisition.* Oxford: Oxford University Press.
- Dörnyei, Z., & Skehan, Z. (2003). Individual differences in second language learning. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589-630). Oxford: Blackwell Publishing Ltd.
- ETS. (2017). *The official guide to the TOEFL® test*. New York:McGraw-Hill.

- Educational Testing Service. (2023). *TOEFL iBT Scores.* Retrieved August 18, 2023, from https://www.ets.org/toefl/test-takers/ibt/scores.html
- Educational Testing Service. (2023). *TOEFL iBT reading section.* Retrieved September 10, from https://www.ets.org/toefl/test-takers/ibt/about/content/reading.html
- Educational Testing Service. (2024). *TOEFL iBT reading section.* Retrieved July 1, from https://www.ets.org/toefl/test-takers/ibt/about/content/reading.html
- Esfandiari, M. R., Riasati, M. J., Vaezian, H., & Rahimi, F. (2018). A quantitative analysis of TOEFL iBT using an interpretive model of test validity. *Language Testing in Asia* *8*(1), 1-13.
- Fahim, M., Bagherkazemi, M., & Alemi, M. (2010). The relationship between test takers' multiple intelligences and their performance on the reading sections of TOEFL and IELTS. Broad Research in Artificial Intelligence and Neuroscience, 1 (3), 0-14
- Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, *3*(4), 70–85.
- Forster, K. I. (1976). Accessing the mental lexicon. In F. Wales & E. Walker (Eds). *New approaches to language mechanisms* (p. 257-287). Amsterdam: North Holland.
- Fox, J. (2004). Test decisions over time: tracking validity. *Language Testing*, *21*(4), 437–465.
- Fries, C. C. (1945). Teaching and learning English as a foreign language. Ann Arbor: University of Michigan Press.
- Fries, C. C. (1963). Linguistics and reading. New York: Holt, Rinehart and Winston.
- Fries, C. C. (1972). Learning to read English as part of the oral approach. In K. Croft (Ed.), Reading on English as a second language: For teachers and teacher-trainers (pp.168- 173). Cambridge: Winthrop Publishers.
- Gebhard, J. (2000). *Teaching English as a foreign or second language: A teacher self – development and methodology guide*. United States of America: The University of Michigan Press
- Golder, K., Reeder, K., & Fleming, S. (2009). Determination of appropriate IELTS band score for admission into a program at a Canadian post-secondary polytechnic institution. IELTS Research Reports, 10, 69–94.
- Goodman, K. S. (1967). Reading: A psycholinguistic game. *Journal of the Reading Specialist, 6*(1), 126-135.
- Gough, P. B. (1984). Word recognition. In Pearson P. D. (Ed.), *Handbook of reading research* (pp. 225–254). New York: Longman.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.
- Grandt, J. (1987). Characteristics of examinees who leave questions unanswered on the GRE general test rights-only scoring (ETS Research Report 87- 83). Princton, NJ: Educational Testing Service.
- Haladyna, T. (2022). Creating multiple-choice questions for testing student learning. *International Journal of Assessment in Education, 9*, 6-18.
- Hammad, E. A. (2021). Palestinian EFL university students' problems with the reading sections of the TOEFL internet-based test and the revised TOEFL paper-delivered Test. *Arab World English Journal, 12* (3), 51-65. DOI:
- Harris, T. L., & Hodges, R. E. (Eds.). (1995). The literacy dictionary: The vocabulary of reading and writing. Newark, DE: International reading Association.

- Hughes, A. (2003). Testing for language teachers. Cambridge: Cambridge University Press.
- Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT®test scores and their use in informing policy in a United Kingdom university setting (TOEFL iBT Research Report No. 30). Princeton, NJ: Educational Testing Service.
- Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. IELTS Research Reports, *2*, 52–63.
- Hill, Y. Z., & Liu, O. L. (2012). Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT Reading performance? (TOEFL Research Report No. 18). Princeton, NJ: Educational Testing Service.
- Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning, 59*(4), 825-865.
- Khonamri, F., & Kojidi, M. E. (2011). Metacognitive awareness and comprehension monitoring in reading ability of Iranian EFL learners. *PROFILE, 13*(2), 99-111.
- Krashen, S. (1982). *Principles and practice in second language acquistion*. Oxford: Pergamon Press
- Kubat, U. (2018). Identifying the individual differences among students during learning and teaching process by science teachers. *International Journal of Research in Educational and Science, 4*(1), 30-38. DOI:10.21890/ijres.369746
- Lee, J. Y. (2011). *Second language reading topic familiarity and test score: Test-taking strategies for multiple-choice comprehension questions*. [PhD Dissertation, University of Iowa].
- Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, *1*(4), 366–389.
- Li, Y. (2018) A comparison of TOEFL iBT and IELTS reading tests. *Open Journal of Social Sciences, 6*(8), 283-309. https://doi:
- Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL ® iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education, 24*(3), 235-255.
- Liu, O, L., Schedl, M., Malloy, J., & Kong, N. (2009). Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning. TOEFL iBT research report. Princeton New Jersey, Educational Testing Service.
- Maizarah, M. (2019). Analysis of the students' common difficulties in TOEFL

reading comprehension at the islamic university of indragiri tembilahan.
EJI (English Journal of Indragiri) : Studies in Edzarnisucation, Literature, and Linguistics, 3(2), 99–106. https://doi.org/10.32520/eji.v3i2.561

- Manganello, M. (2011). *Correlations in the new TOEFL era: An investigation of the statistical relationships between IBT scores, placement test performance, and academic success of international students at Iowa State University* [doctoral dissertation, Iowa State University].
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-790
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: Macmillan Publishing Company; American Council on Education.
- Miller, G. (2017). Metacognitive awareness and reading strategy use: Investigating the intermediate level ESL students' awareness of metacognitive reading strategies. [Master Thesis, St. Cloud State University]

- Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In: Kolers, P. A., Wrolstad, M. E., Bouma, H. (eds) *Processing of visible language* (pp. 259-268). Nato Conference Series, 13. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-0994-9_15
- Ng, J. N. (2007). *Test of English as a foreign language (TOEFL): Good indicator for student success at community college?* [Mater Thesis, Oregon State University].
- Nguyen, T. T. T. (2012). *The Impact of background knowledge and time constraint on reading comprehension of Vietnamese learners of English as a second Language.* [Master Thesis, Southern Illinois University/ Carbondale].
- O'Dwyer, J., Kantarcioglu, E., & Thomas, C. (2018). An investigation of the predictive validity of the TOEFL iBT® Test at an English-medium university in turkey. (TOEFL iBT Research Report No. 83). Princeton, NJ: Educational Testing Service.
- Ovilia, R. (2018). The relationship of topic familiarity and listening comprehension. *Proceedings of the Sixth International Conference of English Language and Teaching (ICOELT 1028) - Advances in Social Science, Education and Humanities Research, 276,* 182-186. https://dx.doi.org/10.2991/icoelt-18.2019.29
- Othman, J., & Vanathas, C. (2017). Topic familiarity and its influence on listening comprehension. *The English Teacher, XXXIV*, 19-32
- Panahi, A. (2014). Threats to validity: construct-irrelevant variances contributing to performance under-representation on Graduate Record Exam (GRE). *Journal of Education & Human Development, 3(*1), 327-346.
- Pekkarinen, T. (2015). Gender differences in behavior under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization, 115,* 94–110.
- Phillips, D. (2015). *Longman preparation course for the TOEFL iBT test*. Ney York: Pearson Education, Inc.
- Priebe, S.J., Keenan, J. M., & Miller, A.C. (2012). How prior knowledge affects word identification and comprehension. *Read and Writing, 25,* 131-149.
- Putlack, M. A., Poirier, S., & Jacobs, A. C. (2020). *Decoding the TOEFL ® iBT: Actual test, reading 2*. Korea: Darakwon.
- Ramos, I, & Lambating, J. (1996). Risk taking: Differences and educational opportunity. *School Science and Mathematics, 96* (2), 94-98
- Rivers, W. M. (1968). *Teaching foreign language skills*. Chicago: University of Chicago Press.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.
- Rogers, B. (2007). *The complete guide to the TOEFL test: iBT edition*. Boston: Thomson Heinle.
- Sadighi, F., & Zare, S. (2006). Is listening comprehension influenced by the background knowledge of the learners? A case study of iranian EFL learners. *The Journal of Linguistics 1*(3), 110-126.
- Schmidt-Rinehart, B. (1994). The Effects of topic familiarity on second language listening comprehension. *The Modern Language Journal 78* (2), 179-198
- Siqueira, M. T., Gonçalves, J. P., Mendonça, V. S., Kobayasi, R., Arantes-Costa, F. M., Tempski, P. Z., & Martins, M. A. (2020). Relationship between metacognitive awareness and motivation to learn in medical students. *BMC Medical Education, 20*(1).
- Sun, M. (2021). Validity and fairness of TOEFL iBT reading test. *Learning & Education, 10*(8), 141-142.

- Sutarsyah, C. (2009). The use of schemata in reading comprehension: A case of learners' reading problems. *Jurnal Ilmu Pendidikan, 16* (2), 67-78
- Syarif, H. (2018). Lexical density vs grammatical intricacy: How are they related? *Advances in Social Science, Education and Humanities Research, 276,* 16-22
- Takala, S., & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323–340.
- Toker, D. (2019). Topic familiarity matters: A critical analysis of TOEFL iBT reading section. *TESL-EJ: The Electronic Journal for Teaching English as a Second Language, 23*(1), 1-9.
- Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review, 8*(1), 1-11
- Vu, L. T., & Vu, P. H. (2013). Is the TOEFL score a reliable indicator of international graduate students' academic achievement in american higher education? *International Journal on Studies in English Language and Literature (IJSELL), 1*(1), 11-19.
- Wait, I., & Gressel, J. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education, 98*(4), 389–398.
- Woodrow, L. (2006). Academic success of international postgraduate education students and the role of english proficiency. *University of Sydney Papers in TESOL, 1,* 51–70.
- Young, D. J. (1991). Creating a low-anxiety classroom environment: What does the language anxiety research suggest? *Modern Language Journal, 75*(4), 426-437.
- Zalha, F.B., Alfiatunnur, A., & Kamil, C. A. T. (2020). Strategies in dealing with the reading section of 'TOEFL prediction': A case of Aceh EFL learners. *Indonesian Journal of English Education, 7*(2), 159-171.
- Zarnis, Y. (2020). Analysis of English education department students' difficulties

in reading *comprehension text of TOEFL* [Diploma, UIN SMH BANTEN].
http://repository.uinbanten.ac.id/5553/

- Zhao, X., & Zhu, L. (2012). Schema theory and college English reading teaching. *English Language Teaching, (5*)1, 111-117.