A META-ANALYTIC INVESTIGATION OF THE PREDICTIVE VALIDITY OF THE TEST OF ENGLISH AS A FOREIGN LANGUAGE (TOEFL) SCORES ON GPA

A Dissertation

by

MAHMOUD EZZAT ABUNAWAS

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Lynn M. Burlbaw
Committee Members,	Homer Tolson
	Robert M. Capraro
	Myeongsun Yoon
Head of Department,	Yeping Li

December 2014

Major Subject: Curriculum and Instruction

ABSTRACT

Research conducted on the relationship between international students' TOEFL scores and academic performance has produced contradictory results. To examine this relationship, a meta-analysis was conducted on extant studies centered on international students' TOEFL scores and academic performance as measured through GPA. The meta-analysis included 47 independent effect size values generated from 40 studies. The results of this meta-analysis yielded a statistically significant and positive relationship between international students' TOEFL scores and their GPA.

As the variation across effect size values was substantial, a moderator analysis was conducted to detect potential variables contributing to this variation. The moderator variables examined in this meta-analysis included: (a) publication status, (b) graduate level, (c) test version, and (d) research setting. Of these four moderators, only the research setting was found to be a potential moderator. Studies conducted outside the U.S. were found to have higher effect size values than those in the U.S. Implications and recommendations were also discussed.

DEDICATION

To My Family

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank Dr. Lynn Burlbaw for his guidance, support, and understanding. I would never have been able to finish my degree without his guidance. He prepared me to grow intellectually and professionally. Thank you, Dr. Burlbaw, for giving me the courage to explore and try, try and fail, try again and succeed. For everything you have done for me, Dr. Burlbaw, I thank you.

I would like also to express my deepest gratitude to my dissertation committee members for their valuable advice and input. In particular, I owe special thanks to Dr. Tolson, whose expertise and knowledge inspired me at every stage of my degree and who spent tremendous effort and time to help me successfully complete my research. I would like thank Dr. Capraro whose insightful suggestions drove me to refine my ideas and improve my research. I would like to thank Dr. Yoon whose classes helped me in my research.

Heartfelt thanks to my family, my parents and brothers, for their sacrifices and everything they have done for me to reach this stage. I would like also to thank my wife and son for their encouragement and patience. Finally, I would like to thank everyone who helped me complete this dissertation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

In this era of globalization, there are approximately 4.1 million students attending universities outside their country of origin (The Organization for Economic Co-operation and Development [OECD], 2012). Currently, the majority of these students are enrolled in English-speaking countries and predominately in the U.S. In fact, during the 2012/13 academic year there were more than 819,644 international students in the U.S. (Institute of International Education, IIE Open Doors, 2013). This suggests almost one in every four of all international students are attending a university in the U.S.

The increasing number of international students has led to a renewed interest in understanding these students' academic performance (Daller & Phelan, 2013; Ren & Hagedorn, 2012). Given the high volume of these students, universities face a major challenge in identifying students who not only meet admission requirements but who also are likely to succeed in completing their academic program. In order to understand the academic performance of these students and inform admission policy for U.S. universities, researchers have examined the predictors related to the academic performance of these students (Annor, 2010; Nelson, Van Nelson, & Malone, 2004; Pitigoi-Aron, King, & Chambers, 2011; Seaver, 2012; Stoynoff, 1997).

Standardized tests are frequently used by universities and researchers as predictors of international students' academic performance. The most recognized of these tests in U.S. universities is the Test of English as a Foreign Language (TOEFL).

The TOEFL test is used by more than 9,000 academic institutions in 130 countries, with the majority located in the U.S. (Educational Testing Service [ETS], 2011a).

Due to the widespread use of the TOEFL, there is a great need for research on the predictive validity of this test. Simner (1999) pointed out this need by stating "The evidence needed to support the TOEFL as a screening device is evidence in favor of predictive validity" (p. 287). Bachman and Palmer (1996) defined predictive validity in the context of language testing as "the extent to which the given assessment predicts candidates' future performance in the target language use domain (p. 46). However, the predictive validity of tests, such as the TOEFL, is usually established by correlating test scores with a measure of past academic performance (e.g., GPA).

Evidence for predictive validity is especially important for indicating how well a test is "useful" for the purpose it was designed. This author subscribes to the belief that predictive validity is "the most important property" of the "practical value" of any test (Schmidt & Hunter, 1998, p. 262). The Standards for Educational and Psychological Testing (1999) highlighted the importance of predictive validity as an essential type of evidence needed to establish test validity (The Standards for Educational and Psychological Testing, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Purpose Statement

There is a growing body of research on the relationship between international students' TOEFL scores and academic performance. However, systematic examinations of the predictive validity of TOEFL scores are rare. There is, therefore, a need for a

meta-synthesis of studies examining the relationship between students' TOEFL scores and academic performance. To extend the understanding of the relationship between international students' TOEFL scores and academic performance, the purpose of this meta-analysis is the integration of existing research centered on students' TOEFL scores and academic performance as measured through GPA.

Research Questions

In this study, the researcher addressed the following research questions:

1. How valid are TOEFL scores in predicting the academic performance of international students as measured by GPA?
2. Do effect sizes vary across studies?
3. Which of the following factors moderate the predictive validity of the TOEFL-GPA relationship: publication status (Published vs. unpublished), degree level (graduate vs. undergraduate), test version (internet-based test (iBT) vs. computer-based test (cBT) vs. paper-based test (pBT), and study setting (U.S. vs. International)?

Rationale for the Study

The researcher in the current study, sought to conduct a meta-analysis that would produce a precise estimate of the predictive validity of the TOEFL scores. As it would provide a valuable contribution to admission decisions about international students, this estimate is of vital interest to differing stakeholders such as students, faculty, and admission officers.

The large number of studies on the relationship between TOEFL scores and academic performance makes it necessary to integrate the results of these studies in a

systematic way. Meta-analysis would be an appropriate method to integrate these results. As discussed above, the results of existing studies concerning the correlation between TOEFL and academic performance appear contradictory. These contradictory results have raised the debate about the validity of TOEFL scores for predicting academic performance. Meta-analysis, therefore, provides a powerful tool for dealing systematically with this variability in research results.

Furthermore, studies of different quality and research design make it a challenge to combine and compare their results. For example, some studies have small sample sizes or other methodological flaws limiting statistical power to detect an accurate predictor-criterion correlation. In contrast to other research methods, meta-analysis addresses this issue by integrating the results of studies with different quality and research designs.

Since 1964, the TOEFL has undergone considerable changes. In the past decade, these changes have included the launching of new versions of the TOEFL, the computer based test (cBT) and the internet based test, (iBT). There is, therefore, a need for a meta-analysis of studies recently conducted on the predictive validity of the TOEFL. As previous meta-analyses were conducted on the old version of the TOEFL (paper-based test), there has been no meta-analysis on the predictive validity of newer TOEFL version (i.e., iBT test).

Previously conducted meta-analyses used inadequate methods of meta-analysis leading to question the validity of their findings. To enhance the validity and generalizability of the current meta-analysis, the researcher, in this study, employed a

variety of statistical techniques neglected in previous meta-analyses. These techniques included the application of fixed and random- effect models; testing for heterogeneity of effect size values; and addressing publication bias.

The TOEFL Test

In 1964, the National Council on the Testing of English as a Foreign Language was formed to develop a test measuring the language skills of speakers of languages other than English, especially for those students applying for admittance into English-medium universities. Beginning in 1965, The TOEFL test has been administered by both the College Board and ETS (ETS, 2011a). The following is a discussion on the development of the TOEFL test which evolved along with developments in second language research.

From a test-purpose perspective, the TOEFL test is described as a language proficiency test measuring the test takers' ability (competence and performance) to understand, read, recognize, and produce standard American Academic English, corresponding to the four language skills: listening, reading, writing, and speaking (ETS, 2011a). By definition, the notion of competence refers to language knowledge while performance is language use (Richards, 2011). Language proficiency tests measure the general ability in the target language and are not tied to a specific purpose, syllabus, curriculum, or material (Brown, 2005). Therefore, the TOEFL test is not an achievement measuring test takers' mastery of specified material.

From a test-reference perspective, the TOEFL test is a norm-referenced test designed to determine test takers' ability in relation to other test takers' ability (Brown &

Hudson, 2002). Thus, the TOEFL test is not a criterion- reference test determining test takers' ability level relative to pre-specified standards. In this respect, the TOEFL test scores are used to compare students with each other to specify who can successfully function in the language of academic contexts.

Norm-referenced tests allow for examinees' scores to be scaled on a normal distribution of abilities where most scores are distributed at the middle of the distribution while extreme scores (lower and higher scores) are on either end of this distribution (Brown & Hudson, 2002). Scores from norm-referenced tests are usually reported as ranks using percentiles and measures of central tendency, including the mean and standard deviation. Scores are typically ranked using three percentiles: 25% of scores are distributed at the bottom 25^{th} percentile, 50% at the middle 50^{th} percentile, and 75% at the upper 75^{th} percentile.

The TOEFL test is also identified as a standardized test in which scores are comparable. The test can be standardized by making the different forms of the test have equivalent content; administrated in similar conditions; and scored according to well-specified standards. Therefore, the reliability and validity of standardized tests are well established through rigorous empirical investigation. Standardized test items are also subjected to several pilot-testing and revisions (Tan & Michel, 2011).

The Paper-Based Test (pBT)

The first version of the TOEFL test was the paper-based test (pBT). This version was based on the structural theory of language which, influenced by behaviorist approaches, views language as a set of discrete elements (Lado, 1961). The test was

initially developed to measure students' mastery of the main language elements including sounds, structures, and vocabulary. The structural theory of language is reflected in both the test structure and content.

The test contains three main sections: listening comprehension, grammar and written expression, and reading. (ETS, 2011a) For content, test items are represented as a series of stimulus eliciting students' responses and formatted as traditional multiple-choice items with four options. This approach is also referred to as discrete-point since test items are independent of each other. The items target the linguistic accuracy rather than the fluency and measures students' ability to produce grammatically accurate element of the language (ETS, 2011a). Additional features of the test are provided in Table 1.

Table 1

Features for Versions of the TOEFL Test

	TOEFL test version		
Feature	pBT	cBT	iBT
Release date	1964	1998	2005
Language theory	Structural	Structural	Communicative
Structure	Listening comprehension; grammar and written expression; and reading	Listening comprehension; grammar and written expression; and reading	Listening, speaking, reading, and writing
Question format	Multiple choice	Multiple choice	Integrated items
Time limit	2 and ½ hours	3 and ½ hours	4 hours
Score range	310-677	0-300	0-120

The Computer-Based Test (cBT)

As the label indicates, the computer-based test (cBT) is the computer-version of the TOEFL used from 1998 to 2005. Although it has the same structure and content of the pBT, this version has different psychometric properties (See Table 1). The main aspect distinguishing this version is that the cBT is an adaptive test. Using computer software, test items are adapted to the level of students' language ability. That is, students do not take the same test and each test is unique to students' ability.

Specifically, items are presented to students based on responses to previous items. Correct responses to items are followed by items of greater difficulty while wrong responses are followed by items of lesser difficulty (ETS, 2011a).

The Internet-Based Test (iBT)

The latest version of the test, introduced in 2005, is the internet-based test (iBT). This version was designed based on new language theories and methods, namely, the communicative approach in which language is viewed as a vehicle for communication (Bachman & Palmer, 1996). The test measures what the takers can do with the language more than what they know about the language. The test structure is composed of four sections: listening, speaking, reading, and writing (See Table 1).

Test items are designed as tasks requiring students to act in authentic academic language situations, such as reading a college-level text or listening to a classroom lecture (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). Unlike previous versions, language skills are viewed as integrated rather than discrete skills. The iBT test, therefore, utilizes integrated tasks in which multiple skills are presented together. In these tasks, for example, students are asked to read a passage about an academic topic and listen to a lecture about the same topic. The students are then asked to use the language to act in these academic contexts such as giving a spoken or written summary of what they read and listened to. The focus of the iBT is more on productive language skills than receptive skills (ETS, 2011a).

Statistical properties for the three versions of the TOEFL test are presented in Table 2. Taken together, TOEFL test versions seem to have high reliability.

Table 2

Statistics for Versions of the TOEFL Test

	TOEFL test version		
Statistic	pBT	cBT	iBT
Mean	524[a]	214[c]	81[e]
SD	65[a]	47[c]	20[e]
Reported Reliability	.95[b]	.95[d]	.94[f]
SEM	13.9[b]	10.8[d]	5.64[f]

[a] Based on 2010 test scores operational data (ETS, 2011b). [b] Based on 1995-1996 test scores operational data (ETS, 1997). [c] Based on 2005-2006 test scores operational data (ETS, 2007). [d] Based on 1998-1999 test scores operational data (ETS, 2001). [e] Based on 2013 test scores operational data (ETS, 2014). [f] Based on 2007 test scores operational data (ETS, 2011c).

CHAPTER II

LITERATURE REVIEW

The important issue of understanding the relationship between international students' TOEFL scores and academic performance was identified in the previous chapter. Researchers have provided evidence concerning the interpretation and use of students' TOEFL scores (Chapelle, Enright, & Jamieson, 2008; Sawaki & Nissan, 2009; Wang, Eignor, & Enright, 2008). In this evidence, researchers have focused on the predictive validity of these scores on measures of academic performance (e.g., Ayers & Quattlebaum, 1992; Cho & Bridgeman, 2012; Van Nelson, Nelson, & Malone, 2004; Vinke & Jochems, 1993; Wait & Gressel, 2009). However, results of this research have been contradictory and therefore its use in admission decisions is controversial (Jameison, Jones, Kirsch, Mosenthal, & Taylor, 2000; Vu & Vu, 2013). Therefore, a review of literature related to international students' TOEFL scores and academic performance is presented in this chapter.

A number of researchers have found a positive association between students' TOEFL scores and academic performance (e.g., Cho & Bridgeman, 2012; Torres & Zeidler, 2002; Wait & Gressel, 2009). As a result, TOEFL may prove a promising predictor of academic performance. However, other investigators have found low associations between students' TOEFL scores and academic performance (e.g., Neal, 1998; Ng, 2007; Vinke & Jochems, 1993; Xu 1991; Yule & Hoffman, 1990). The results of these studies put into question the use of TOEFL scores in admission decisions.

As part of his influential work on the relationship between language proficiency and academic success, Graham (1987) presented a review of studies conducted on the relationship between students' TOEFL scores and academic performance. Only half of the studies included in the review had positive results on the predictive validity of TOEFL scores. The author concluded that the research provided inconclusive evidence as to whether the TOEFL should be used as an indicator of students' academic performance.

Cho and Bridgeman (2012) have collected evidence on the predictive validity of international students' TOEFL scores on academic performance. Cho and Bridgeman's study of 2594 undergraduate and graduate students from different fields of studies at 10 U.S. universities has been one of the most extensive studies to date. The authors found a small association (r = .16 for graduate students, and r = .18 for undergraduate students) between students' TOEFL scores and GPA. These results suggest that around 3% of the variance in the GPA is explained by students' TOEFL scores.

In searching for more evidence of the predictive validity of students' TOEFL scores on the academic performance of these students, Cho and Bridgeman (2012) used contingency tables in which each student under study was placed in one of three subgroups based on TOEFL scores (top 25%, middle 50%, and bottom 25%) and GPA. The results of the subgroups analysis showed students with high GPAs tend to have high TOEFL scores and students with low GPAs tend to have low TOEFL scores. Cho and Bridgeman (2012) concluded that although the correlations found in their study were

small, the results of the subgroups analysis supported the predictive validity of students' TOEFL scores on academic performance.

Wongtrirat (2010) conducted a meta-analysis of correlation values between students' TOEFL scores and academic performance of international students in U.S. universities. Academic performance for students was measured using both GPA and course completion. Including 20 studies between 1997 and 2007, Wongtrirat found a correlation of .187 between students' TOEFL scores and GPA. Although Wongtrirat (2010) included both published and unpublished studies, the meta-analysis was not comprehensive and included only two studies on students' TOEFL scores and course completion.

The meta-analysis by Wongtrirat (2010) included studies in which only the older version of the TOEFL test (pBT) was examined. In addition, Wongtrirat's analysis did not account for variation in effect sizes nor examine moderating factors. In analyzing the difference in correlations between undergraduate and graduate levels, the author focused only on the mean difference between the two groups of studies without using the appropriate statistical procedure of testing homogeneity of effect sizes. Due to these limitations, this meta-analysis failed to give a complete picture of the relationship between student's TOEFL scores and academic performance. In an earlier synthesis of the relationship between students' TOEFL scores and first-year GPA, Yan (1994) determined a correlation value of .3 in 27 studies published between 1964-1994. (As cited by Yan, 1995). In addition to the limitations of the meta-analysis by Wongtrirat (2010), Yan (1994) used only the first-year GPA as a measure of academic performance.

Challenges to Validity

The discussion on predictive validity should not proceed without presenting a review of challenges associated with predictive validity research. Validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (The Standards for Educational and Psychological Testing, 1999, p. 9). Establishing the validity argument for tests requires a continuous process of collecting evidence to support interpretations and uses of test scores. Recurring themes in validity include, therefore, the "interpretations" and "uses" of score inferences. Hence, "validity is not a property of the test or assessment", (Messick, 1995, p.741), rather the interpretation and uses drawn from test scores.

Applying this validity argument on the current study suggests test scores are valid predictors for specific criterion describing a certain population (Brown & Coughlin, 2007). In this respect, rather than academic performance, international students' TOEFL scores are more appropriate predictors of their academic English proficiency (Palmer and Woodford, 1978). This view is echoed by the ETS statement that students' TOEFL scores should not be used "as the sole criterion for admission or as a predictor of academic success" (ETS, 1994, p. 8). This section presents a review of literature associated with challenges for predictive validity, criterion challenge, and the link between language proficiency and academic performance.

Predictive Validity Challenge

Predictive validity studies in education often yield low or moderate validity values (Cohen, 1988). Researchers, in these studies, usually conceptualize predictor and

criterion variables in a unique manner that is determined by the constructs assumed to underlie variables of interest. In language testing, the same constructs may be operationally defined differently (Bachman, 2004). For example, language proficiency may be defined as a competence or communicative performance (Shohamy, 1996).

The validity coefficient depends on the relevance of criterion variables (Sawaki & Nissan, 2009). These criteria are related to the uses and interpretations of the test. For example, first-year GPA is a common criterion variable for academic performance. However, this variable might not always be a relevant criterion (Sawaki & Nissan, 2009). Especially when GPA does not fully represent the academic performance construct. More discussion on selecting relevant measures follows.

In addition to being relevant, the predictor and criterion variables must be reliable (Liao, 2004). The criterion unreliability, which would certainly underestimate the validity coefficient, might be due to different sources of measurement error. Most variables especially in educational fields would have measurement error. Measurement theories such as classical test theory, which were founded on the notion of measurement error, fundamentally aim to investigate the sources of this error. According to classical test theory, the observed test score is composted of the true score plus a random error where the true score is the score that would be obtained over repeated times of the measurement. Hence, reliability is concerned with the degree to which the test measures the same construct consistently. Sources of measurement error include personal factors such as guessing as well as testing conditions such as testing time and scoring

procedures. This measurement error could obscure the true test score of and subsequently the correlation between the predictors and criterion variables.

Selecting appropriate variables becomes more challenging when these variables are contaminated with other construct-irrelevant factors (Messick, 1989). The criterion variables (e.g., GPA) are usually related to other variables (e.g., age and gender) that might affect the validity coefficients and limit the generalizability of study results. While the effect of these variables might not be possible to be removed from the study, these variables can be dealt with by either including them in the investigation or using other statistical methods such ANCOVA. The problem of contamination might lead to another concern in prediction studies known as differential prediction which results from the possibility of having different results for different groups of participants based on factors such as: age, gender, or language background (Young & Kobrin, 2001).

Uncontrolled variables might also affect the predictive validity coefficient such as the time lag between the predictor and the criterion. Researchers confirmed the notion that the correlation between students' language proficiency and first-year GPA is higher than that of cumulative GPA (Light, Xu, & Mossop, 1987; Elder, 1993). These results are consistent with research findings on the predictive validity of other admission tests, especially the GRE, and suggest these tests are more predictive for students' first-year GPA (Kuncel & Hezlett, 2007).

Moreover, because of the effect of time lag, students' TOEFL test scores are valid only for two years from the test date. To illustrate, the correlation between students' TOEFL scores and GPA would be stronger for recent scores than older ones.

For example, Yule and Hoffman (1990) found the correlations between students' TOEFL scores and graduate GPA at 12 months (r = .24) is stronger than at 18 (r = .03) or 24 months (r = .15).

The predictive validity coefficient may be also attenuated due to range restriction (Hunter & Schmidt, 2004). This suggests that the samples from which the coefficient is estimated are normally restricted to only those students admitted to university rather than the full population of applicants. Therefore, the range restriction reduces the variability of scores and consequently the validity coefficients are underestimated. Since the validity coefficient is based on sample variance, having little variance in scores will ultimately lead to a weaker correlation (Goodwin & Leech, 2006).

In the case of students' TOEFL scores, the sample is also restricted by the fact that selected candidates are more likely to have higher TOEFL scores while those with low TOEFL scores are not usually admitted (Cho & Bridgeman, 2012). Range restriction might also lead to having homogeneous groups of scores. Based on TOEFL scores, for instance, students are usually classified into one of three proficiency level types: high, medium, and low. Researchers indicated that correlation between language proficiency and academic performance is stronger for the low proficiency levels than for the high levels (Elder, 1993). For example, in his study on the relationship between English proficiency and academic performance, Woodrow (2006) found English language has a more effect on international students' achievement in the low proficiency levels than in the high ones.

Having homogeneous groups of scores might lead to having nonlinear relationship between the predictor (e.g., TOEFL) and the criterion (e.g., GPA) (Goodwin & Leech, 2006). This nonlinear relationship might result from dichotomizing or categorizing continuous variables. As a result, correlation analysis may not be an appropriate method of measuring these nonlinear variables. As another example, in a comparison of IELTS and TOEFL scores as predictors of GPA, Hill, Storch, & Lynch (1999) found the correlation between IELTS scores and GPA was strong. However, the correlation was weak for TOEFL scores. Hill, Storch, & Lynch (1999) believed this weak correlation between the TOEFL and GPA might be due to the curvilinear relationship between the two variables.

Range restriction could also provide an explanation for the small correlation between admission measures and academic performance at more selective universities. These universities generally require high TOEFL scores and GPAs from students. Due to this lack of variability, the correlation coefficient between these measures is expected to be small regardless of the actual relationship (Goodwin & Leech, 2006).

Researchers recommend to statistically correcting for the attenuated correlation that result from different sources of measurements errors including unreliability (Hunter & Schmidt, 2004) and range restrictions (Glass & Hopkins, 1996). However, applying the correction for measurement error resulting from either criterion unreliability or range restriction requires additional data which are not always available in studies such as the reliability estimate for the criterion and the predictor variables.

Sample size is also an important factor influencing the predictor-criterion correlation. Using small samples might lead to an unstable estimate of the validity coefficient (Goodwin & Leech, 2006). There are no specific guidelines on the optimal sample size for designing predictive validity studies in educational research; however, some researchers recommend a sample size above 200 (Schmidt, Hunter, & Urry, 1976). More details on the effect of the sample size are detailed in the methods chapter of this dissertation.

Criterion Challenge

As international students have different social, cultural, and educational systems, prediction of their academic performance is a complex process. Predicting the success of these students using only test scores is inefficient as these scores do not fully account for the complexity of academic performance. Researchers suggest that academic performance is multivariate in nature and is better captured by a variety of factors (Kuncel, 2003; O'Connor & Paunonen, 2007).

The complexity of academic performance along with challenges in predictive validity research discussed above leads to what can be called "the criterion problem". This problem centers on identifying a valid and appropriate criterion measure for academic performance. The cornerstone of designing a predictive validity study is to determine the nature of the criterion to be predicted. Having a well-specified definition of what represents academic performance will certainly lead to more accurate predictive results (Kuncel, 2003).

Researchers examined the predictive validity of language tests have traditionally relied on one academic criterion measure, the grade point average (GPA). Of all the different types of GPA (e.g., a semester GPA, year GPA, and cumulative GPA), first-year GPA is the most frequently used measure of academic performance in predictive studies (Noble & Sawyer, 2002; Zahner, Ramsaran & Steedle, 2012). Using GPA, however, as an indicator of academic performance has several limitations. For example, grading standards vary by academic institutions, departments and even between courses (Lei, Bassiri, & Schulz, 2001). Therefore, GPA does not account for the variability within and across universities.

GPA is also limited due to the restriction of range in grades. The effect of range restriction becomes exacerbated due to grade inflation (Lei, Bassiri, & Schulz, 2001). Regardless of these limitations, GPA continues to be the most commonly used criterion measure of students' academic performance. Kuncel (2001) stated even those who criticize GPA use the measure as a primary indicator of students' academic performance. Indeed, GPA is a more convenient and an objective measure. Other measures of academic performance are either subjective such as evaluations or are limited indicators of students' performance such as retention.

Researchers have confirmed GPA has a high internal reliability (Bacon & Bean, 2006). Kuncel (2003) documented from previous studies a GPA reliability estimate between .80 -.84. GPA also has the advantage of not only being used as an indicator of academic performance but also as a predictor of academic success (Kuncel, Hezlett, & Ones, 2001) and employment success (Strenze, 2007). In addition, researchers suggests

GPA is an effective indicator of non-cognitive measures of academic performance such persistence, commitment, and learning strategies (Allen 1999; Fenster, Markus, Wiedemann, Brackett, & Fernandez, 2001) as well as effort and motivation (Bacon & Bean, 2006). Finally, GPA can effectively predict whether students will continue enrollment in their degrees (Murtaugh, Burns, & Schuster, 1999).

Linking Language Proficiency and Academic Performance Challenge

A fundamental assumption in predictive validity research is the predictor and the criterion must be closely related; that is, they must measure the same construct. It is, therefore, important to delineate how students' language proficiency is related to their academic achievement. Researchers suggest English language proficiency is associated with the academic performance of non-native English speaking students (Cho & Bridgeman 2012; Stoynoff, 1997). Researchers who advocate this view argue that having English language skills enables non-native English students to meet the requirements of their degrees and improve their academic performance. A visual illustration of how language proficiency as measured by the TOEFL might be related to academic achievement is depicted in Figure 1.

Contrarily, some researchers believe language proficiency is not associated with academic performance. They suggest that even students who pass the TOEFL face challenges with academic language (Ren & Hagedorn, 2012). In addition, they argue that international students with limited English skills are able to tackle the academic requirements of their degrees, regardless of proficiency (Stoynoff, 1997).

A related concern is whether a standardized test score such as the TOEFL can be used as a proof of the student's language ability to function in academic settings. The answer to this question depends on the degree to which the TOEFL measures the same language abilities that are needed in academic English contexts. This is a question of the construct validity and indicates the extent to which the test measures what it is supposed to measure. The construct validity of the TOEFL was the subject of extensive investigation that provided different types of evidence that the TOEFL does measure academic English proficiency (Chapelle, Enright, & Jamieson, 2008; Sawaki, Stricker, & Oranje, 2008).

Much of the debate on the relationship between language proficiency and academic achievement has arisen from the lack of a theoretical framework explaining how language proficiency is related to academic performance (Cummins, 1984). A simple depiction of this framework is displayed in Figure 1. As demonstrated in this figure, although it is related to academic performance, language proficiency does not explain all of the variance in academic performance.

The researcher, through a review of the literature, found that language proficiency explained less than 10 % of students' academic performance. The remaining variance may be explained by several factors, such as motivation, gender, language background, and country of origin. These factors could also moderate the relationship between language proficiency and academic performance. Therefore, these factors might lead to inconsistencies in the results of studies on the predictive validity of language tests (Cho & Bridgeman, 2012).

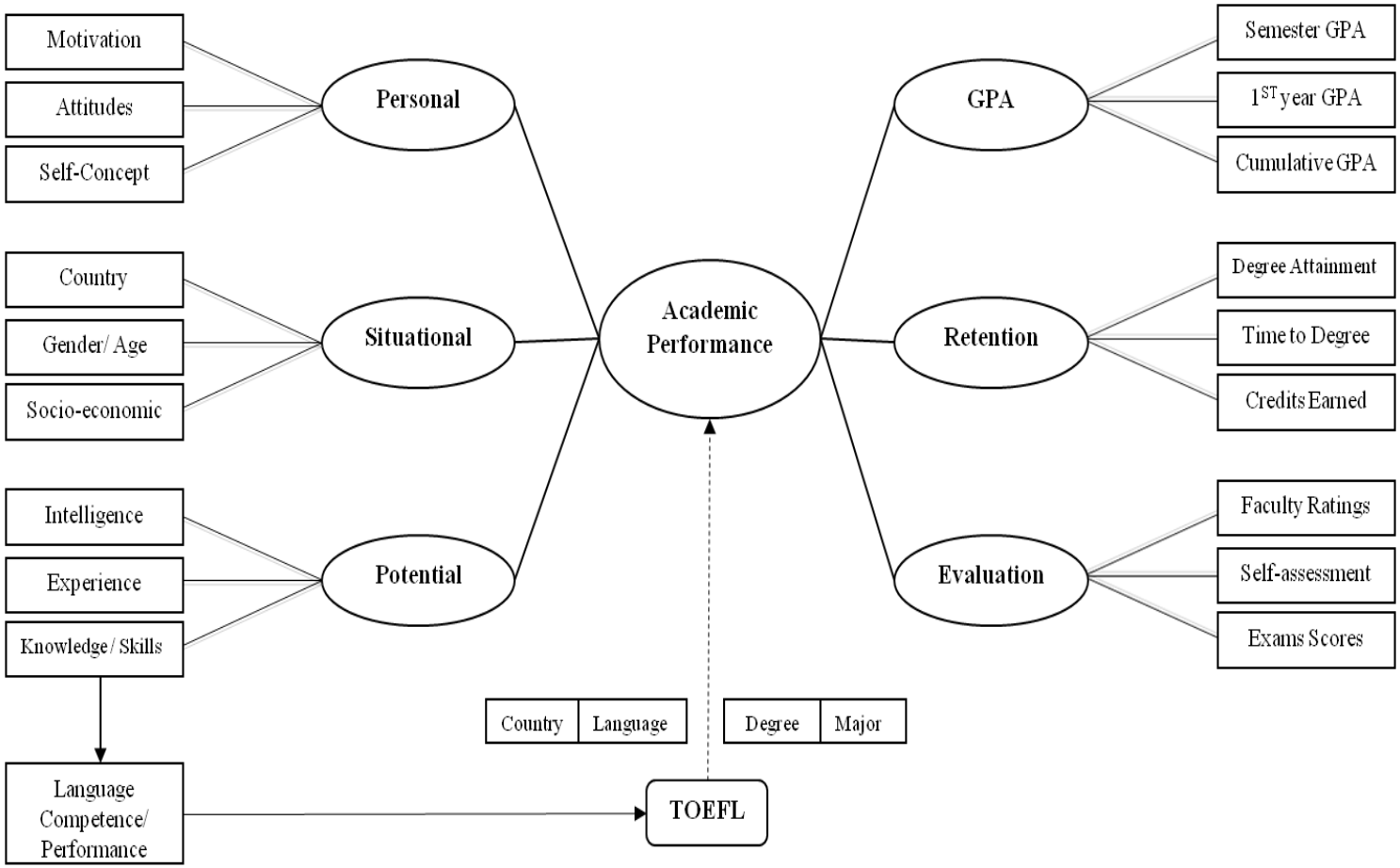

Figure 1. Theoretical Model of the Relationship between Language Proficiency and Academic Performance

The discussion on the predictive validity challenges shows how the predictor-criterion relationship is an important issue that should be considered when studying the relationship between language test scores and academic performance. This issue was further studied by Hartnett and Willingham (1980) who concluded the predictor could still be inappropriate as the criterion, GPA, does not fully represent the construct of interest, academic performance.

An Interpretive Argument-Based Approach to Validity

The validity framework of this study drives from Messick's view of validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (Messick ,1995, p.741). Rather than separate types of construct, content, and criterion, validity is a unitary concept in which all types of evidence are linked. Messick's notion of validity also introduced a new perspective on validation as a process involving judgment.

In addition to test scores interpretations and uses, Messick's framework added test-based "actions". These actions are related to the consequences of test scores on those who are affected by these scores including test takers, test users, schools, policy makers, and society. By including test actions, Messick's validity view laid the foundation for the social dimension of test scores. The consequences of test scores-based actions include test philosophy, values, fairness, ethics, and the impact of test scores on decision-making (McNamara & Roever, 2006).

A notable conceptualization of Messick's validity framework adopted in this study is the argument-based approach to validity. Based on Messick's (1989) notion of validity, the argument-based approach was proposed by Kane (2006) and drawn from Toulmin's (1958) argument model. This approach was elaborated and investigated by several language testing researchers (e.g., Bachman, 1990; Chapelle, Enright, & Jamieson, 2008).

The traditional validity frameworks involve two dimensions of validity evidence: theoretical rationale and empirical evidence (Chapelle et al., 2008). The argument-based approach to validity, on the other hand, includes a third dimension: logical evaluation (Kane, 2001). Validity is viewed as not only a process of accumulating different types of theoretical and empirical evidence supporting the validity of test scores but also as a logical evaluation of the interpretations based on test scores. Kane (2006) suggested two stages for building the validity argument for a test, starting with the argument-developing stage in which the intended inferences and claims underlying the inferences are proposed. The second stage, the argument-evaluation stage, identifies the types of evidence backing up the claims and the threats to the validity.

A validity argument for the TOEFL test scores is presented in Figure 2. First, the cycle framework indicates the validation process is a continuous process conducted at any stage of the test development and administration. The figure highlights the main inferences explicitly made on the test scores: test scores-based interpretations are appropriate for the intended purpose if the test items are adequately representative of the target construct and consistently measure what is supposed to be measured. In this

model, test performance is associated with other tests measuring the same construct and provides a predictor of future performance on the same construct. The bottom of the figure shows an example of the decision-making process based on these inferences.

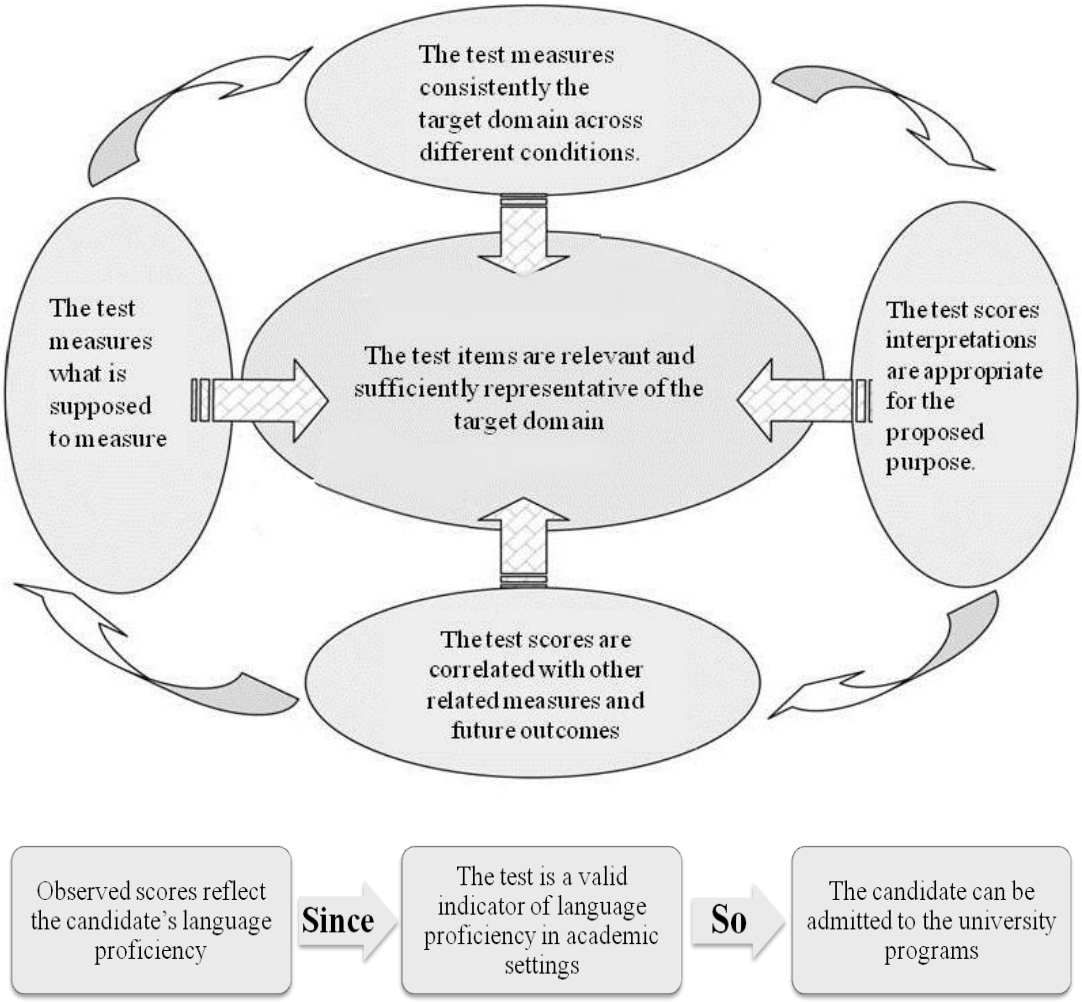

Figure 2. An Interpretive Argument-Based Approach to Validity

Test Score Interpretation and Use

As previously discussed in this chapter, validity concerns test scores-based interpretations and uses (Messick, 1989). The inferences drawn from the test should be interpreted with caution as they could invalidate scores regardless of the test's validity (Bachman, 2005). Setting rigid cut-off scores to determine who will be admitted might lead to misclassification of students. For example, misclassification of students based on their test scores may result in admitting unqualified students or denying qualified students (Xi, 2008).

To adjust for misclassification, a variety of factors should be considered when specifying cut-scores for an admission test, including: nature of the program, admission requirements, number of students, and profile of applicants (Des Brisay, 1994). For example, university programs and departments requiring more English proficiency might have higher cut-off scores. Therefore, the admission process should consider not only the quantity of the test scores but also the overall quality of the student. Using TOEFL as an example, universities should consider the TOEFL section scores in addition to the total score in their admission process. Each TOEFL section (reading, listening, speaking, and writing) provides different contributions to the total TOEFL score (ETS, 2005).

Universities and departments, therefore, should consider setting their own cut-off scores. However, as the ETS, that administers the TOEFL test, do not recommend specific cut-off scores, individual universities and departments should set these cut-off scores according to their own academic standards. The author, through search of the admission requirements of selected universities, revealed that the TOEFL cut-off scores

ranged from 61-100 with the majority of these universities requiring between 70 and 80. These cut-off scores also vary within individual institutions and across degree levels. Some universities also set a range of TOEFL scores while others do not set a specific score. The TOEFL cut-off score requirement in most universities is close to the mean TOEFL iBT score which, based on the most recent TOEFL scores summary (2014) is 81 with a standard deviation of 20.

TOEFL scores do not classify students in terms of language proficiency level (low, intermediate, and high). Rather, as discussed previously, TOEFL scores reflect a student's level in comparison to other test takers. However, the ETS has published a range of level equivalents for each section score. As a result, a TOEFL score of 80 suggests a student is likely to have an intermediate to high language proficiency level.

The practice of setting test cut-off scores is usually arbitrary and not supported by research (Chalhoub-Deville & Turner, 2000). Setting appropriate cut-off scores is a challenge as they are usually revised and modified according to university policy and standards. The process of setting cut-off scores should therefore go through standard setting which involves a shared judgment of all people involved in the process, including: (a) policy makers, (b) students, (c) faculty and (d) admission staff (Wylie & Tannenbaum, 2006). This process supports the conclusion, "There is no absolute, unequivocal cut score. There is no single correct or true score" (Wylie & Tannenbaum, 2006, p. 2).

As was previously discussed in this chapter, a test score might not be an accurate representation of the student's true score as there is always a source of measurement

error. This error might be due to a variety to factors related to the test taker or the test administration. Therefore, when interpreting a test score, test users are recommended to consider this error by calculating the standard error of measurement (SEM). This measurement provides an estimate of certain ranges within which the true test scores should fall (ETS, 2005). It is noteworthy that the larger the SEM, the less accurate the estimate of the candidate true score.

The Predictive Validity of the TOEFL

This section provides a literature review of the research on the TOEFL predictive validity. Since its launch in 1966, numerous studies have been conducted to examine the predictive validity of the TOEFL (Ayers & Quattlebaum, 1992; Cho & Bridgeman, 2012; Van Nelson, Nelson, & Malone, 2004; Vinke & Jochems, 1993). The researchers, in these studies study, have examined TOEFL predictive validity by correlating students' TOEFL scores with measures of academic performance (Fu, 2012; Johnson, 1988; Neal, 1998; Ng, 2007; Wait & Gressel, 2009). However, this research has yielded conflicting results concerning the predictive validity of the TOEFL test (Jameison et al., 2000).

A number of researchers found the TOEFL test is a potential predictor of academic performance (e.g., Ayers & Peters1977; Cho & Bridgeman, 2012; Stoynoff, 1997; Torres & Zeidler, 2002; Wait & Gressel, 2009). Yet, these researchers have found low correlation values. On the other hand, some researchers found no significant association between students' TOEFL scores and academic performance (Neal, 1998; Ng, 2007; Vinke & Jochems, 1993; Xu 1991; Yule & Hoffman, 1990).

Regardless, Simner (1998) questioned the validity of using students' TOEFL scores as an admission measure. He refuted the claims that the poor results of the TOEFL predictive validity might be due to inadequate sample size or other methodological flaws such as range restriction. Furthermore, he claimed these results reflect a weak correlation between the TOEFL and academic performance. Finally, Simner (1998) provided evidence that the academic performance of students with low TOEFL scores is comparable to that of students with high TOEFL scores and even to native English students. Rather than targeting the TOEFL test itself, Simner (1998) explicitly criticized how universities use the TOEFL in making admission decisions.

Van Nelson, Nelson, and Malone (2004) investigated potential predictors of academic performance for 866 international students in an American university from 1987-2002. In addition to TOEFL scores, the researchers modeled the following variables: age, gender, country, language, academic major, GPA of first 9 hours, and admission status. Academic performance was measured by completion of degree and graduate GPA. As a justification for the use of TOEFL as a predictor of academic performance; the authors correlated TOEFL scores with GRE scores. Having found a correlation of .50 between the GRE and the TOEFL, the authors concluded TOEFL scores could be used as a predictor of academic performance.

In their study, Van Nelson, Nelson, and Malone (2004) dichotomized the final GPA into two categories: (a) below 3.5 and (b) above 3.5. The author found students' TOEFL scores are a better predictor of degree completion. However, when combined with other predictors, such as undergraduate GPA, TOEFL scores were found to have

predictive power for GPA. Although the results were not supportive of students' TOEFL scores as predictor of academic performance, the researchers recommended using TOEFL scores in combination with other variables, especially undergraduate GPA. The results are limited due to range restriction of the high TOEFL scores.

In a recent study by Fu (2012), the predictive validity of prior GPA, SAT, GRE, and TOEFL was conducted on the first-year GPA for undergraduate and graduate international students in a U.S. public university. Based on the correlational analysis, the TOEFL scores had a significant correlation with academic performance. However, in the multiple regression analysis, the TOEFL scores were not a significant predictor when other measures were included in the model. These results coincide with Schmidt (1991) who tested a multiple regression model for a set of predictors taken from students. These predictors included, (a) freshmen GPA (b) high school rating, (c) ESL grade, and (d) SAT score. Schmidt (1991) found including the SAT-verbal had the same predictive power as TOEFL scores.

In contrast, Cho (2012) found TOEFL scores add to the prediction of students' GPA beyond the SAT and the GRE-verbal. In addition, having found a high correlation between the TOEFL and GRE- verbal, Ayers and Peters (1977) concluded that students' TOEFL scores and GRE verbal scores could be combined as an effective predictor for academic performance.

TOEFL and Academic Performance by Major

Wait and Gressel (2009) explored the relationship between students' TOEFL scores and academic performance for 6,516 international engineering students enrolled

at an American university in the United Arab Emirates. Indicators of academic performance included GPA for courses in engineering and in humanities, Comprehensive Assessment Examination (CAE), and graduation rate. The model also included data on the different university colleges and engineering majors. The results suggested a relationship between TOEFL scores and academic performance.

Specifically, Wait and Gressel (2009) found TOEFL scores were a better predictor for arts and sciences majors rather than engineering. Similarly, Hughey and Hinson (1993) found that the association varies by major; the correlation values were statistically significantly higher for humanities and sciences majors than business majors. Interestingly, Hughey and Hinson (1993) also found that the association was lowest for undeclared majors, suggesting the relationship between students' TOEFL scores and GPA may be affected by other factors.

Recently, Cho & Bridgeman (2012) revealed that the association between TOEFL scores and GPA is higher for students in business, humanities, and social science majors than those students in science and engineering majors. Likewise, Light, Xu, and Mossop (1987), found a small but statistically significant association between TOEFL scores and academic performance of graduate students which varied by academic major. The relationship was stronger for social science and education majors than for science and business majors. These results suggest differences in the association between TOEFL scores and academic performance by academic discipline may be due to differences in language demands across disciplines.

TOEFL Sections

Researchers suggest that students' TOEFL scores prediction of academic performance may vary by test section. A study by Perry (1988) found the receptive skills have more predictive validity than oral skills. These results were comparable to Johnson's (1988) study who found that the structure, vocabulary and reading sections of the TOEFL test are more highly associated with academic performance than the listening section. Al-Ansari and Al-Musawi (2003) found the TOEFL vocabulary and reading section was a more effective predictor of undergraduate GPA than the listening, structure and written sections.

Researchers, in the above studies, examined the paper-based TOEFL (pBT) subsections. Few researchers examined the new TOEFL iBT subsections which include the four sections of listening, speaking, reading and writing. Fu (2012) conducted a comparison between the predictive validity of the two test versions and found that the two versions are similar in their correlations with GPA with the exception of the listening section. The TOEFL iBT listening sections was a more effective predictor of GPA while the TOEFL pBT was not. These results reflect the improvements on the new version of the TOEFL iBT which includes more tasks that are closer to academic contexts such as lectures.

In his study, Fu (2012) also indicated that TOEFL correlation between TOEFL with first-year GPA varies by sections and across degree levels. For the undergraduate level, the reading section had the highest correlation with GPA while for graduate level the writing had the highest correlation. Furthermore, the speaking section was a better

predictor for graduate than undergraduate student's performance. The implication of these findings is that the language skills needed for graduate levels are different than that of undergraduate levels. The productive skills (speaking and writing) are more important for a graduate student whose course work requires more oral presentations, communication with faculty, and writing skills.

Demographic Variables

Few researchers examined whether the correlation between TOEFL and academic performance might vary by demographic variables such as gender, country and native language. Hughey and Hinson (1993) found the correlation between the TOEFL and GPA varies by native language; the correlation was stronger for students whose native languages are related to English such as European languages rather than language that is different form English such as Asian languages.

In their study, Wait and Gressel (2009) revealed that TOEFL had more effect on the academic performance of female students than male students. This is in contrast to the results of Hughey and Hinson's (1993) study in which they found that although gender might be an important factor in determining academic performance, the correlation between TOEFL scores and cumulative GPA was equally low and significant for both males and females.

TOEFL Cut-off Scores

Light, Xu, and Mossop (1987) found that students with low TOEFL scores had higher GPAs than those with higher TOEFL score. This finding could be due, according to the authors, to other unexplained factors such as students, background and motivation.

Similarly, Lo (2002) found that students who meet the TOEFL pBT cut-off scores (above 550) tend to have a lower GPA (below 3.0) than students who did not meet the TOEFL pBT cut-off scores (below 550).

Ng (2007) found no significant differences in the academic performance of international students based on their TOEFL scores. Low TOEFL achievers tend to have equal academic performance with high TOEFL achievers. Stoynoff (1997) determined that TOEFL scores were correlated with GPAs; however, through interviews, he also determined that students scoring at or above the cut-off scores still face challenges in their studies.

In contrast, Johnson (1988) who found a positive correlation ($r = .36$, $p < .01$) between TOEFL and GPA, concluded that students with low TOEFL (below 500) have lower GPAs while students with high TOEFL scores (above 500) have high GPAs. These results are comparable to the study by Hu's (1991) that showed that students who scored more than (575) had a higher GPA than students who scores less than (575). Using 348 foreign students from 50 U.S. universities, Messner and Liu (1995) examined the validity of the TOEFL cut-off scores. The authors discovered that students with TOEFL scores higher than 550 had higher GPAs. Similarly, Cho and Bridgeman (2012) found that students with top 25% GPA are more likely to be with top 25% TOEFL scores than from bottom 25% TOEFL scores.

TOEFL and Other Indicators of Academic Performance

A number of researchers examined the correlation between the TOEFL with a variety of academic performance indicators other than GPA such as placement tests,

retention, and self -assessments. Using data on 503 international students who were enrolled at Iowa state university between 2009 and 2011, Manganello (2011) assessed the relationship between the TOEFL iBT total score as well as the subscores and the Iowa state university's English Placement Test scores (EPT).

Although the correlation between the TOEFL and the EPT sections was statistically significant and ranged from (r = .31 to.41), the author concluded that that these results are not enough to establish a relationship for the TOEFL iBT to be used instead of the EPT in making placement decisions. This conclusion was made based on the small amount of shared variance (ranged from 11% to 17%) between two variables (TOEFL and EPT) that are supposed to measure the same construct (Academic English).

Manganello (2011) also correlated the TOEFL scores with grades on four English composition classes and found a statistically significant correlation (r = .291). Unlike the older versions of the TOEFL whose correlation with similar course grades was less than .20, the author concluded that this correlation indicates that the TOEFL is a valid predictor of course grades. In addition, Manganello (2011) found the correlations between the TOEFL and course grades were significantly higher for the writing (r = .257) and the speaking (r = .174) sections of the test than for the reading (r =.004) and the listening (r = .042) sections.

Kwai (2009) investigated the factors related to the retention of 454 international undergraduate students in two public university systems between 2006 and 2007. Retention was defined in terms of students' progressive re-enrollment. The variables included country of citizenship, financial sponsorship, admission status (freshmen or

transfer), gender, TOEFL, GPA, number of credit hours, and on-campus integration. The author used the TOEFL instead of high school grades because of the differences of the high school systems across countries. Although it was not correlated with retention, the TOEFL was found to be indirectly related to students' retention through it is correlation with GPA.

A number of researchers examined the correlation between the TOEFL with other admission tests such as GRE. Stricker (2004) compared the performance of native and nonnative speakers of English on both the computer- based TOEFL and GRE. Although the TOEFL test is specificity designed for the nonnative speakers of English, the author justified including the native speaker group for the sake of comparison. For the native speakers group (n =168), the correlations were .61, .44, and .39 for GRE verbal, quantitative, and analytical sections respectively. For nonnative speakers group (n =3489), the correlations were .64, .34, and .53 for GRE verbal, quantitative, and analytical sections respectively. These moderate to high correlations are consistent with other studies conducted on the TOEFL-GRE correlations in which the correlation between TOEFL and GRE-verbal sections were higher than those for GRE-quantitative sections (Pennock-Román, 2002; Yule & Hoffman, 1990).

Neal (1998) examined the correlation of the GRE as well as the TOEFL scores with the graduate GPA (GGPA) of 47 native Indians and Chines graduate students in the sciences and engineering programs at the Rose-Hulman Institute of Technology. The study revealed that while there was a significant correlation between the GRE-quantitative and GGPA, no significant correlation was found for either the GRE-verbal

or TOEFL with GGPA. These results were similar to Ayers and Quattlebaum (1992) who found a significant correlation between the TOEFL and GRE but not between the TOEFL and GPA.

Other researchers investigated the correlation between TOEFL and number of credit hours (e.g., Stoynoff, 1997). In his study, Johnson (1988) identified a high correlation between TOEFL and number of credits (r = .80). On the contrary, Light, Xu, and Mossop's (1987) results revealed a small but significant correlation between TOEFL and the number of credit hours completed (r = .19). These findings were comparable to an earlier study by Bower et al., (1971) who also found a small significant correlation between TOEFL and number of credit hours (r = .18).

While most researchers examined the TOEFL predictive validity by correlating the TOEFL with objective measures such as GPA, few researchers tried to use other subjective measures such as evaluations. Vu and Vu (2013) examined the correlation between the TOEFL and the GPA of 464 international graduate students at a U.S. university and found a significant but weak correlation (r = .117). In addition, the authors designed a survey about students' perceptions of the TOEFL. It was found that students believe that their TOEFL scores are related to their academic performance. These results were similar to Wang, Eignor, and Enright (2008) who found a correlation between the student's TOEFL scores and self-assessment of their English skills in a form of responses on a questionnaire.

TOEFL and Non-academic Variables

Researchers tried to find if there is a relationship between the TOEFL and non-academic variables that affect the academic performance of international students such as acculturation (de Souza, 2012); adjustment (Gong & Fan, 2006; Wang, 2003); and locus of control (Schmit, 2001). These researchers made the claim that English proficiency could be associated indirectly with international students' academic performance by helping them adapt to the new culture and consequently improving their academic achievement. For example, Zhang (2012) found that students with higher TOEFL scores are more likely to have less acculturative stress and a better academic performance.

TOEFL is used by several universities in predicting the performance of international teaching assistants' (ITA) (Farnsworth, 2013). Few researchers examined the correlation of the TOLEF test scores with a variety of ITAs' performance indicators including: student evaluations of the ITA (Witt, 2010); TA assignment outcomes (Xi, 2008); TA examination (Kamara 1994); or recommendation to teach (Yule & Hoffman 1990). Xi (2008) found that TOEFL iBT speaking scores were a significant predictor of the TA assignment outcomes. Also, Yule& Hoffman (1990) discovered that students with higher TOEFL scores are more likely to have recommendation to teach.

Summary

In this chapter, the researcher provided a literature review about the TOEFL predictive validity, and a review of critical issues related to the predictive validity of the TOEFL. In completing this review, the reader was exposed to the pertinent research on

the relationship between international students' TOEFL scores and academic performance. The results of the studies reviewed in this chapter were inconsistent. This uncertain picture about the TOEFL predictive validity necessitates conducting meta-analysis, which will be the focus of the remaining chapters. In the next chapter, the reader is provided with the methodology the researcher employed to conduct this meta-analysis.

CHAPTER III

METHODS

To examine the relationship between the TOEFL and academic performance for international students as measured by grade point average (GPA), the researcher implemented a meta-analysis methodology. In this chapter, the procedures for conducting the meta-analysis are presented, including: (a) the case for meta-analysis, (b) scope of the search, (c) search strategy and sample selection, and (d) analytic framework and procedures. The researcher discusses in the following section the case for using meta-analysis to examine the relationship between international students' TOEFL scores and academic performance as measured by GPA.

The Case for Meta-Analysis

The term meta-analysis was first introduced by Glass (1976) as "the analysis of analyses" (p. 3) and refers to the quantitative method of integrating the results of several related studies in a systematic manner. With the purpose of combining and /or comparing these results, meta-analysis yields a precise estimate of an average effect size summarizing the strength of relationship or the magnitude of difference between variables of interest (Cooper, 2010). If the effect size values drawn from studies are homogeneous, meta-analysis may help identify possible reasons for a common effect. If, on the other hand, effect size values are heterogeneous (i.e., varies from one study to the other), meta-analysis techniques can be utilized to identify possible reasons for heterogeneity.

Meta-Analysis as Synthesis of Knowledge

The use of meta-analysis to synthesize knowledge occurs in different areas of research. A close look at peer-reviewed journals suggests meta-analysis has high citation rates (e.g., Review of Educational Research). As meta-analysis is used to provide results from the analysis of multiple studies, meta-analysis results frequently inform-decision making (Lundahl, Yaffe, & Hobson, 2008). In fact, there is little question as to whether meta-analysis is important to research; there is always a need to cumulate past research to inform current practices and generate future research. Additionally, advances of knowledge occur not only by conducting more experiments and observations but also by synthesizing the results of studies to reach conclusions or make decisions.

Narrative reviews use an interpretive epistemology to synthesize knowledge but are often criticized as being associated with subjective bias. Unlike narrative reviews, meta-analysis shares the same design features of analytic studies, from problem formulation to hypothesis testing. Therefore, meta-analysis is a method rooted in positivistic epistemology. In addition, meta-analysis provides researchers with a set of specified procedures allowing other researchers to replicate the results of a meta-analysis.

By handling a large number of studies, the possibility of missing studies in meta-analysis is much lower than the case with narrative reviews. Even in the case of missing studies, advanced meta-analysis techniques provide solutions for dealing with missing studies. In addition, meta-analysis allows for efficient coding of study characteristics

(e.g., sample size and research design) and enables researchers to examine differences between studies (Garg, Hackam , & Tonelli, 2008).

Meta-analysis is also a more powerful method than other quantitative methods of pooling studies such as vote counting in which the number of studies with statistically significant results are counted and compared with the ones with non-significant results. Since they are based on the significance of results in synthesizing studies, these traditional methods of quantifying results usually do not provide clear-cut answers to research questions. For example, as discussed in the previous chapter, in his review of studies on the relationship between TOEFL and academic achievement, Graham (1987) was not able to reach a conclusion. In contrast to these traditional methods, meta-analysis can be used to tell us about the size of the effect regardless of the significance of the results, which enables researchers to reach conclusions about the effect.

Meta-Analysis as Contributor of New Knowledge

Meta-analysis not only provides a method for synthesizing existing knowledge but also contributes new knowledge to an existing area of research. This is possible as research questions used in individual studies are transferable to meta-analysis studies. In the current meta-analysis study, for example, one question centers on the relationship between international students' TOEFL scores and academic achievement. While this question was examined by a number of researchers in individual studies, the current study used meta-analysis to answer the same question. Thus, results from this meta-analysis contribute new knowledge to the existing research on international students.

Advantages of Meta-Analysis

In addition to synthesis of existing knowledge and contributions to new knowledge, meta-analysis provides certain advantages over individual studies. For example, a meta-analysis provides a platform for studying inconsistencies in a particular field of research. By studying variation of effect size values across studies, the current study seeks to understand inconsistencies in previous individual studies on the relationship between students' TOEFL scores and academic achievement. Furthermore, meta-analysis can be used to test and contribute to current theories. The present study might give valuable insights on the theoretical perspectives of the relationship between language proficiency and academic achievement by examining potential variables that moderate this relationship.

Meta-analysis studies can have stronger statistical power in comparison to individual studies. Studies with small sample sizes might have less power to find significant results than those with large sample sizes which increases the risk to commit Type II error. While there are several factors that affect the significance of the results, the sample size is considered the main factor that determines whether the results will reach statistical significance (Field, 2013). By including a large number of studies that are weighted by their sample sizes, results drawn from meta-analysis are not sensitive to sample size.

In addition to its sensitivity to sample size, statistical significance testing does not inform us about the practical or substantive significance of the results (Gliner, Leech & Morgan, 2002). Studies could yield significant results but with no practical value. The

limitations of significance testing established the need for a precise measure of practical significance, namely the effect size, which is the main product of meta-analysis.

Effect Size

Researchers' efforts to explain effect size have greatly contributed to our understanding of meta-analytic methods. However, there is still a difficulty in identifying a precise meaning of effect size (Kelley & Preacher, 2012). Cohen (1988) defined the concept of effect size as "the degree to which the phenomenon is present in the population" (Cohen, 1988, p. 9-10). Effect size represents a standardized measure of the magnitude of the difference or the strength of relationship between the variables of interest (Cooper, 2010).

Effect size has received a great deal of attention in research (Fidler, Thomason, Cumming, Finch, & Leeman 2004; Thompson, 2008) especially by Cohen (1990) who stated that "the primary product of a research inquiry is one or more measures of effect size, not p values" (p. 1310). In its recent Publication Manual (2010), the American Psychological Association (APA) called for reporting effect size values in studies: "For the reader to appreciate the magnitude or importance of a study's findings it is almost always necessary to include some measure of effect size in the results section" (p. 34). So, researchers are always recommended to use the effect size in reporting and interpretation of the study results (Wilkinson and the Taskforce on Statistical Inference, 1999).

Despite the increasing number of journals that require researchers to report effect size measures (Thompson, 1999), there are still a large number of published journal

articles that do not report these measures (Fritz, Morris, & Richler, 2012). McMillan and Foley (2011) examined how researchers report and interpret effect size in educational research. The authors reviewed 417 articles of quantitative and mixed-methods that were selected from four educational journals as a sample of educational research. They found that 74 % of the articles included effect size measures and only 51% of them had an interpretation of the reported effect size values. McMillan and Foley concluded that although there is an increase in effect size reporting, providing an interpretation with the effect size is very limited in educational research. According to the authors, these results indicate that researchers do not have a sufficient understanding of effect size.

Scope of Search

The current study is a comprehensive meta-analysis of the available studies in which the researcher examined the relationship between TOEFL scores and the academic performance as measured by GPA for international students. The choice between either a restricted or comprehensive search strategy will have direct implications on the outcomes of the meta-analysis. The first choice is to limit the search by restricting the inclusion criteria on type of studies included in the analysis. Although restricting the inclusion criteria to type of study is a more convenient option, this restricted approach is more likely to exclude relevant studies and therefore limit the power of meta-analysis to detect the effect of interest.

The second choice is to employ a comprehensive search strategy capturing a large number of available studies. While this choice generates a more representative sample of studies, enhancing the power of meta-analysis to detect the effect of interest,

this choice is less convenient to researchers as it requires more time and effort to retrieve and manage a large database of studies. Another issue with implementing comprehensive search strategy is included studies usually vary by quality and design, introducing the challenge to combine and compare study results.

In this meta-analysis, the researcher employed an inclusive approach to locate studies on the relationship between international students' TOEFL scores and GPA. The rationale for using this inclusive approach was to gather more complete picture of the research conducted on the topic of interest. By maximizing the number of studies to be included in the analysis, this approach avoids biases associated with excluding relevant studies. This approach also improves the validity of the analysis and increases the power of meta-analysis to detect a stronger relationship between the variables of interest.

The search process consisted of two phases. To identify the depth and the scope of the search, the researcher implemented an initial search phase of the topic of interest. The second phase included the procedures described in the following section.

Inclusion Criteria

Informed by the purpose and the research questions of the study, the inclusion criteria for studies in the meta-analysis were specified. Studies were included in the meta-analysis if they meet the following criteria:

1. Variables of interest: relevant studies must have included a quantified value describing the relationship between international students' TOEFL scores and academic performance as measured by the GPA. The predictor measure must have been the TOEFL scores, including any of the three versions pBT, cBT, and iBT. The

criterion measure must have been GPA, including an accumulative GPA or a more specific type of GPA such as first-year GPA.

2. Target participants: non-English speaking students including both graduate and undergraduate students who took the TOEFL in order to study at English-medium universities. These universities can be located in English speaking countries or universities located in non-English speaking countries in which English is the medium of instruction.
3. Analysis: eligible studies must have included quantitative results and sample sizes allowing the estimation of effect size between variables of interest. As a result, analyses could include any of the following: Pearson's correlation, regression, *t*-test, ANOVA, Chi-Square, direct effect size, or descriptive statistics (means and standard deviations).
4. Publication status: all studies published, unpublished, or otherwise available after 2000 can be included in the analysis. The decision for restricting studies to this date is due to the growing body of research on TOEFL during this period, especially newer studies relevant to the topic under examination. In addition, this period witnessed major developments of the TOEFL, especially introduction of new test versions.
5. Report language: included studies must have been in English. Non-English studies were not included due to practical difficulties in translation.

Search Strategy

"How one searches determines what one finds; and what one finds is the basis of the conclusions of one's integration of studies" (Glass, 1976, p.6). In this quotation, Glass suggests that the search process shapes the outcomes of meta-analysis. The following is a detailed account of the procedures used for searching the relevant literature. Following the inclusion criteria listed above, the researcher conducted an exhaustive search of available research on the relationship between international students' TOEFL scores and academic performance.

To retrieve the maximum number of relevant studies, the search process was iterative and flexible with no restricted rules applied at the initial stage of the search. The researcher initiated the search by identifying major resources used in locating the relevant studies. These resources included electronic as well as printed resources of published and unpublished research.

Published Research

The electronic search was implemented using the major electronic databases: ERIC, ProQuest, Science Direct, EBSCO, SAGE, and other related databases. The researcher identified and searched relevant journals (e.g., Language testing) using full text journal databases (e.g., Wiley, Taylor & Francis, and JSTOR). In addition to the electronic databases, several search engines were used to identify potential studies (e.g., Google Scholar and the ETS website).

Journals only available in print were searched manually. Additionally, a comprehensive library search for relevant books and handbooks was conducted. To

identify further relevant studies, the researcher performed an ancestry search by examining the references of the retrieved articles. Moreover, a search of subsequent citations of the located articles was performed using special citation indexes, such as the Web of Science.

By employing basic and advanced search techniques, electronic search was performed using multiple combinations of the following words and phrases: TOEFL, academic achievement, performance, success, academic outcomes, GPA, admissions, predictive validity, English proficiency, test scores, standardized tests, examination, reliability, university or college international students (graduate, undergraduate).

Unpublished Research

A search for relevant dissertations and theses was conducted through the ProQuest electronic Dissertations and Theses database; repositories of selected U.S. and international libraries; and WorldCat. Also, Conference papers were searched using conference websites pages (e.g., Language Testing Research Colloquium). Requests for relevant articles and ongoing research were made using email, listservs, and other related websites. In addition, authors in the field were contacted using email to ask for the availability of any related studies.

Data Extraction

The data extraction materials were developed and revised throughout the research process and included a coding form and coding guide. To ensure the systematicity of the process, the following procedures were established. In developing the coding form, the first step was to determine the type of data to be extracted.

Following recommended coding procedures (Cooper, 2010; Lipsey & Wilson, 2001), the researcher specified a priori the items that were coded from studies. As a more objective coding procedure, setting a priori coding criteria enhances the quality of meta-analysis as it ensures that data collection is informed by the research purpose and inclusion criteria and not by the coded studies (Cooper, 2010). Nevertheless, there were some types of data that emerged during the pilot testing of the coding form and were added to the coding scheme. As the coding sheet was expanded to include the new items, the previously coded studies were coded again for the new added items.

The next step was to develop the coding form in which the researcher tabulated the actual items that were collected from each study using an electronic spreadsheet (Microsoft Excel, 2010). The coding form was organized according to the following categories: study characteristics, sample characteristics, measures of interest, moderators, and results (See appendix B). The coding guide included details describing the coding items and provided step-by-step instructions for the coding process.

There are two levels of coding; low inference coding requires filling the data directly from the study and high inference coding requires making decisions about the coding items (Cooper, Hedges, &Valentine, 2009). For an illustration of high level coding, in the current study, while the coding sheet included only a code for the type of academic outcome (e.g., GPA) reported in studies, the coding guide provided instructions on the different types of outcome (e.g., final GPA). The coding guide also included details how the outcomes are measured and examples that exhaust most of the possible options for coding this item.

The first major coding category included the codes for study characteristics. Each study was identified by the authors last names and linked to the full text as wells as a complete citation in APA style. In addition to the publication year, the publication status was also recorded as the following: journal article, thesis or dissertation, book/book chapter, technical report, conference paper, unpublished manuscript, or other type of publication.

Since it is used in the estimation of effect size values, the total number of participants in each article (*N*) was recorded. As some studies included more than one variable besides the TOEFL, the coding guide was refined to include instructions on coding only the sample size of the TOEFL variable rather than the total sample size. The coding sheet also included entries for the measures including the predictor (i.e., TOEFL scores) and the outcome variable (GPA). The specific measure of the outcome and how the measure was operationalized in studies was also recorded. For example, for the GPA, the measure could be first-year GPA or cumulative GPA.

For descriptive purposes, other variables of interest were recorded if they were reported in studies, including academic disciplines, and TOEFL section scores. Common demographic variables were also recorded including gender, age, country, or language of the participants. Unfortunately, a small number of studies included separate results for these variables. For example, only five studies included results for academic disciplines and only three studies included results for the TOEFL sections scores. Due to insufficient data, these variables were not included in the analysis.

The moderator category included codes for variables believed to lead to any systematic differences between the effect size values. The degree level was coded as undergraduate or graduate. Other information describing the academic level of students was also recorded such as Master, PhD, community college, professional students, nontraditional students, or other type of students. The test version entry included three codes (pBT, cBT, or iBT). The study setting was coded as either U.S. or international including the name of the country.

The results category included the type of statistical analysis and the actual statistical results. The type of statistical analysis refers to the statistical test that was used to test the relationship between the predictor variable (i.e., TOEFL scores) and the outcome variable (GPA). The statistical test could be one of the following categories: inferential statistics (e.g., correlation, regression, *t*-test, ANOVA), descriptive statistics (e.g., Means, SDs): a measure of effect size (e.g., eta-square), *p*-values, non-parametric tests (e.g., Chi-Square) or any other type of statistical test that can be converted to an effect size.

Along with the statistical tests, the actual results of studies were coded including page number(s) where the results were located. The results could be the numeric value of any of the test statistics mentioned above including: correlation coefficient (*r*), regression coefficient (*B*), *F*-Statistic, *p*-values or other effect size indexes. The coding sheet included the results for the total sample as well as the results for the subsamples if they were reported in studies.

The coding form included an open entry to record any important findings, issues, or notes that come up during the coding process. Any items that did not conform to any of the coding categories were coded as "other" with the accompanying details. The coding sheet also included information about the missing data. For example, in the case of studies that did not refer to the test format (pBT, cBT, or iBT), the entry was coded as "missing".

Multiple Outcomes per Study

Multiple results reported from a single study are not uncommon when extracting data for a meta-analysis. For example, in this search, the researcher located studies that used two types of GPAs to measure academic performance. Reporting multiple effect size values in the same study raises the issue of statistical dependency assumption which requires that effect size values must be statistically independent from each other (Lipsey & Wilson, 2001). Correspondingly, multiple effect size values drawn from the same study violate this assumption of independency.

A common practice to deal with the situation in which multiple outcomes reported in the same study is to combine these outcomes (Borenstein, Hedges, Higgins, & Rothstein, 2009). This option is usually applied when these multiple outcomes are measuring the same construct and estimated from the same sample. For example, if a certain study included two results for both first semester GPA and second semester GPA, these two estimates can be averaged into one outcome. In this meta-analysis, a number of located studies included results for both first-year GPA and cumulative GPA.

The researcher decided to include only the cumulative GPA since it is a more representative measure of academic performance than the first-year GPA.

On the other hand, if the outcomes are not related, they measure different constructs or represent independent subsamples, the conventional solution is to treat these results as separate studies and analyze them accordingly (Borenstein et al., 2009). Therefore, studies with multiple outcomes will produce more than one effect size. For instance, for this meta-analysis, one study reported multiple correlations for the three TOEFL versions (iBT, cBT, pBT). Although they come from the same study, the results do not violate the assumption of dependency since they represent independent subsamples. Therefore, three effect size values were recorded for this single study, one for each test version.

Evaluating the Coding Procedures

To ensure the credibility of the coding process, the researcher applied several procedures to make the coding process "transparent and "replicable" (Card, 2011). Transparency requires describing the details of the coding decisions. Replicability entails that the coding procedures are fully explained so that other researchers can get similar results if they were to repeat the same process. To achieve transparency and replicability of the coding process, as previously explained, the full details of the coding procedures were provided including the coding materials with detailed instructions of what to code and how to code each item.

As coding procedures are susceptible to subjective judgment, the coding sheets were shared with expert colleagues in the field to ask for their input about the content

and the structure of the coding sheet. Furthermore, the researcher performed a pilot testing of the coding materials was a random selection of the articles. By identifying the items that needed to be refined, the researcher revised and improved the coding sheet.

Following the recommended standards of conducting and reporting a high quality meta-analysis such as the Preferred Reporting Items for Systematic Reviews and Meta - Analyses: The PRISMA Statement (Moher, Liberati, Tetzlaff & Altman, 2009), the author gave great care to document the full details of the actual search procedures. In addition, copies of all the articles in PDF files were stored and organized in a separate folder.

Sample Selection

The multiple-stage search process generated a number of results with irrelevant and duplicate items across all databases. The search results were filtered several times to remove these duplicate and irrelevant items. The researcher examined the titles and abstracts of the retrieved articles and identified a number of potential studies for further review. (See appendix C for flow diagram that lists the number of included and excluded studies).

Having identified the population range of available studies, the researcher scrutinized these retrieved studies to check if they fully met the inclusion criteria. At this stage, the researcher conducted a more in-depth screening of each article's full text. This screening process excluded additional studies because they did not meet one or more of the inclusion criteria as follows: A number of studies were not found and most of these studies published outside the U.S. Other studies did not report complete information

about the sample. For example, a study by Fakeye and Ogunsiji (2009) used multiple samples including the TOEFL and other variables. Since the authors reported one correlation for the multiple samples as a combined sample, it was not possible to tell the results for the specific TOEFL sample and therefore was excluded from the current analysis.

Additionally, some studies were eliminated because they did not have enough statistical data. Other studies were removed as they were qualitative studies, or non-empirical studies. The researcher repeated the search process until no new relevant studies were found. The researcher then, identified those studies to be included in the final analysis.

All retrieved articles with the complete codes were tabulated in the spreadsheet with rows representing the studies and columns including the entries for coded data. To perform the analysis, the studies and coded items were imported into the meta-analysis software, Comprehensive Meta-Analysis (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005).

Analysis

In meta-analysis, the unit of analysis is the studies rather than the participants of studies. The analysis was conducted in the following consecutive steps: first, estimating the effect size values from individual studies; second, computing the weighted average for the effect size values obtained in the previous step; third, investigating the variability of the effect size values across studies; and fourth, conducting a moderator analysis on the variables that could contribute to the variation in values.

Analytical Framework

The decision to choose the meta-analysis framework involves two issues. Meta-analysts should first specify the statistical method that will be used to estimate the average effect size. The most common methods for conducting meta-analysis are the Hunter and Schmidt (2004) method and the Hedges and Olkin (1985) method. Second, the models that will be applied in analyzing the data should also be determined. Meta-analysis data are usually analyzed using fixed or random effects models. The following is a discussion of the methods and models that were employed to conduct the current meta-analysis.

The Hunter and Schmidt method generates a weighted average effect size using the raw data whereas the Hedges and Olkin approach calculates the weighted average effect size using transformed data. Field (2001) conducted a comparison of the two methods and found that both of them suffer from biasing the estimated effect size especially when the sample of studies is either small or heterogeneous. According to Field (2001), when there is an effect in the population, the Hedges and Olkin method leads to upward bias in the effect size estimate while the Hunter and Schmidt method results in a downward bias. Both methods can be less biased if the studies are homogeneous which would be an unrealistic assumption in practice as studies are usually varied by sample size, setting, and instruments used (Field, 2001).

Since it requires making several corrections on the individual correlations drawn from studies, the Hunter and Schmidt approach uses corrected correlations rather than the actual ones in estimating the average effect size. Therefore, the Hunter and Schmidt

method is criticized as being an estimate of what the effect size should be under ideal conditions rather than what the effect size actually is (Rosenthal, 1991). Another limitation of using this approach is the data required to apply the corrections are not always readily available, they are either hard to obtain or not reported in all studies. Hunter and Schmidt (2004) provided solutions for these missing data but they are also based on hypothetical assumptions. Therefore, the Hedges and Olkin method was deemed the better approach for the current meta-analysis.

Meta-Analysis Models

The meta-analysis model (fixed or random) needs to be chosen at the beginning of the data analysis stage. Choosing a certain model will affect the meta-analysis outcomes since each model has different assumptions and approaches of estimating and interpreting the effect size. Therefore, based on the nature of the data, each model might produce a different effect size estimate. The decision for applying one of these models is based on the assumptions about the variability of studies and the degree of results generalization.

The fixed effect model assumes that there is one true effect size in the population from which the sample effect size values are drawn. This model also assumes that the variation of effect size values across studies does not reflect a genuine variation between values and that the source of variation is due to random error (Hedges & Vevea, 1998). This error is the only source of variation that is accounted for in estimating the effect size. Therefore, if this source of error was removed, which is practically not possible, all studies would have the same effect size value. In other words, the effect size values in

population studies are homogeneous. Accordingly, using this model, the results are not generalizable to other studies beyond those included in the analysis (Field & Gillett, 2010).

In contrast to the fixed effect mode, the random effect model assumes that the effect size values in the population from which the sample values are drawn vary. Therefore, changing the study features such as the sample might alter the effect size estimate. The variation of the effect size values across studies is a genuine variation and reflects not only a random error but also a systematic deference between studies (Hedges & Vevea, 1998). The studies represent a random sample of the population in which the effect size values are heterogeneous; therefore, the results can be generalized to all other studies representing that population (Field & Gillett, 2010).

In brief, choosing between the two methods will have a direct implication on the statistical results. There is a trade-off between practicality and usefulness in choosing one of the models. The fixed effect model is less complicated (it includes only one error term) but generates a less precise estimate of the true effect size (National Research Council, 1992). In addition, applying the fixed effect model will be at the expense of losing valuable features of meta-analysis such as examining the variability of effect size values and the follow up moderator analyses proposed in this study.

The random effect model is more complicated (i.e., it includes two error terms) but it overcomes the limitations of the fixed effect model mentioned above. Therefore, this meta-analysis was conducted using the random effect model because the researcher sought to examine the variability of effect size values and to generalize results to the

population of interest. The fixed effect model was also estimated to allow a comparison of the two models.

Analysis Procedures

Initial Data Screening

An initial data check was performed in order to detect any coding errors or inaccuracies in the data. The coding errors could mislead the data analysis and interpretation of the results. Therefore, to detect such errors, descriptive statistics such as the mean and the standard deviation were examined. Characteristics of the sample were also presented such as the number of studies and the total number of sample sizes. In addition, using frequency distributions, graphs, and figures, data were screened to check for normality and any extreme outliers in the data.

Calculating Effect Size Values

The first step of the data analysis is to extract the effect size values from the individual studies. Since most of the studies included in the analysis reported correlational analysis, the Pearson product-moment correlation coefficient (referred to as Pearson's *r*) is the effect size index that was used in the analysis. If the measure of effect size was not directly reported in a study, then an effect size was obtained from available data reported. For example, some studies included statistical tests other than correlation (e.g., *t*-test, *F*-test); the effect size was computed by converting the reported statistics into the correlation coefficient (*r*) following conventional conversion procedures.

To enable the results of the various studies to be combined and compared, the obtained correlations were then converted into a common scale, namely, Fisher's *r* to *z*.

As a standardized measure, this effect size index has the advantage of making the results of different studies comparable (Lipsey & Wilson, 2001). The rationale for this transformation is that the sampling distributions of correlations are not normally distributed, thus violating the assumption of normality required in combining correlations from different samples (Rosenthal, 1991). The purpose of Fisher's *r* to *z* transformation, therefore, is to normalize these distributions and to stabilize its variance. The Fisher's *r* to *z* transformation was applied using the CMA software and obtained by:

$$Z_r = .5 \ln\left[\frac{1+r}{1-r}\right]$$

The CMA software preform the analysis using these transformed *z* effect size values denoted as (ES) rather than the correlations. For practical difficulties in Fisher's z interpretations, since (*r*) is a more common and preferred index over the transformed z, Fisher's z was converted back to the correlation (*r*) to facilitate the interpretation of results. Converting *z* back to *r* can be obtained in the CMA software by the formula:

$$r = \frac{e^{2ES_{Zr}} - 1}{e^{2ES_{Zr}} + 1}$$

The transformed values (ES) along with the studies and the sample size were listed in the CMA software spreadsheet utilized in the calculations of the average effect size.

Calculating the Weighted Average Effect Size

Having obtained the effect size values (ES) from each study, the next step was to estimate the total average. Rather than simply pooling the average, the effect size values were first weighted by their sample sizes. As studies are dissimilar and have different

sample sizes, weighting effect size values with their corresponding sample sizes produces more precise and less biased estimates. The average effect size was computed by giving more weight to studies with more sample sizes and less weight to the studies with less sample size. Therefore, studies with more power (large sample size) have more contribution in the total effect size than studies with less power.

The effect size values were weighted by multiplying each effect size by the inverse of its variance (called the within study variance) using the following procedures (Cooper et al., 2009). These procedures was performed using the CMA software.

First, the variance (v_i) of each effect size (ES) is calculated as:

$$v_i = \frac{1}{N-3}$$

The weight (w_i) of each study is the inverse of this variance:

$$w_i = \frac{1}{v_i}$$

The weighted average effect size ($\overline{ES}$) is then calculated as:

$$\overline{ES} = \frac{\sum(w_i ES_i)}{\sum w_i}$$

Each individual effect size (ES_i) is multiplied by its weight (w_i) then summed { $\sum(w_i ES_i)$ } and divided by the sum of the weights ($\sum w_i$)

The average effect size is the indicator that was used in answering the first question about the relationship between TOEFL scores and academic achievement as measured by GPA. A positive effect size indicates that there is a relationship between

the predictor (i.e., TOEFL) and the outcome (i.e., GPA). The interpretation of the effect size follow the criteria suggested by Cohen (1988): $r = .10$ (small effect), $r = .30$ (medium effect), and $r = .50$ (large effect). Although they are useful, these criteria should be interpreted with caution since their interpretation depends on the nature of study and the research context (Ferguson, 2009). Thompson (2002) recommended interpreting effect size values in relation to existing research findings rather than using specific criteria.

Random Model

The average effect size presented in the previous section was estimated using the fixed effect (FE) model. To apply the random effect (RE) model, in addition to the within study variance (v_i) used in the fixed effect model, the RE model incorporates a between study variance (v_θ). This new source of variance will have implications on estimating the weight:

$$W = \frac{1}{v_i + v_\theta}$$

So, the inverse of the variance is calculated by the CMA software using both the within and between study variances

The between study variance (v_θ) is estimated by the CMA software as (Hedges & Olkin, 1985):

$$v_\theta = \frac{\left[\sum w_i (ES_i - \overline{ES})^2\right] - (df)}{\sum w_i - (\sum w_i^2 / \sum w_i)}$$

This formula estimates the weighted sum of squared deviations of each effect size values from the weighted average effect size where the *df* is the number of studies (*k*) minus 1. More information about this formula will be explained in the section on heterogeneity of effect size values.

Confidence Intervals

As effect size provides only a point estimate of the effect, a more efficient procedure is to report effect size values with their confidence intervals (Kelley & Preacher, 2012). According to Glass & Hopkins (1996), confidence intervals are "a range of possible values, so defined that there can be high confidence that the 'true' values, the parameter, lies within this range" (p. 261). While the effect size is a measure of strength, the confidence intervals are measures of precision. The current edition of the APA's Publication Manual (2010) calls for reporting confidence intervals specifically on effect size values: "Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size" (APA, 2010, p. 34).

Using confidence intervals, the sample is used to estimate the extent to which the population value is likely to fall within certain levels of confidence (usually 95%). The confidence interval (CI) for the average effect size is calculated by the CMA software as the following:

$$CI = \overline{ES} \ \ 1.96Se$$

Where Se is the standard error of the average effect size and equals the square root of the variance:

$$Se = \sqrt{\frac{1}{\sum Wi}}$$

The obtained 95% CI values tell us that 95% /100 such interval would include the population average effect size. The wider the CIs, the more variable and less precise is the effect size.

The interpretation of effect size values can be facilitated by examining the position and width of their corresponding confidence intervals. Effect size values and confidence intervals for individual studies as well as the total average effect size are usually outlined using effective and informative graphs such as forest plots which is garneted by the CMA software. In addition to highlighting the variability of individual and average effect size values (indicating the precision of the effect size value), confidence intervals (CIs) can also be used in making inferences about the statistical significance of the average effect size (Lipsey & Wilson, 2001). If the CIs does not include zero, the effect size value will be statistically significant. If the CI includes zero, it will be concluded that the effect size is not statistically different from zero.

In interpreting the findings, the statistical significance criterion should not be used as an indication to existence or the absence of the effect. In the words of Sagan (1997), "Absence of evidence is not evidence of absence!" (p.200). Non-significant results might be attributable to the fact that the study does not have a sufficient power to detect the effect of interest. As discussed previously, the power to find statistically significant results is influenced by other factors especially the sample size. (Field, 2013)

Heterogeneity of Effect Size Values

The second question of this study was to examine the variation of effect size values; whether these values are consistent or varied across studies. Meta-analysis

contributes to research not only by quantifying an overall effect size, but also investigates the nature of effect size values across studies. Unfortunately, a large number of meta-analyses, including the previous one conducted on this topic, sought to find out whether there was an effect or not. To answer the second question, testing the variability of effect size values was conducted using the *Q* statistic (the Cochran Chi-Square test for heterogeneity) which is estimated in the CMA software as (Hedges & Olkin, 1985):

$$Q = \sum w_i (ES_i - \overline{ES})^2 = \sum (w \times ES^2) - \frac{\left[\sum (w \times ES)\right]^2}{\sum w}$$

This *Q* test estimates the weighted sum of squared deviations of each effect size from the weighted average effect size. It is noteworthy that this test is the numerator of the formula of the between study variance (ν_θ) that is used in estimating the random model as presented previously. This makes sense as the larger the *Q* value, the more variable the effect size values, and the wider the confidence intervals around those values. The obtained value of this *Q* statistic is evaluated using the corresponding critical value from the chi-square distribution with $k - 1$ degrees of freedom where *k* is the number of studies. If the *Q* statistic value exceeds this critical value, the null hypothesis that the effect size values are homogenous will be rejected and it will be concluded that the variation between effect size values is larger than what would be expected by chance (sampling error) alone (Lipsey & Wilson, 2001).

The *Q* test suffers from the same limitations of statistical significant tests. Specifically the test is not robust to number of studies with the potential for non-significant Q test results (Borenstein et al., 2009). The sample size is a determining

factor for most statistical significance tests (Field, 2013), including the *Q* test. Therefore, researchers may choose to use other tests robust to sample size. In addition, the *Q* test indicates only whether the effect size values are homogeneous in terms of statistical significance without specifying the degree to which these effect size values vary. To address these concerns, Higgins and Thompson (2002) proposed the index (I^2) that is robust to number of studies.

$$I^2 = \frac{Q - \mathrm{df}}{Q}$$

The index (I^2) is defined as the ratio of the between study variance to the total variance and estimates the amount of the true variance between effect size values. Large (I^2) indicates heterogeneity and its value will be interpreted according to the following criteria: I^2 = 75%: large heterogeneity; 50%: moderate heterogeneity; and 25%: low heterogeneity (Higgins, Thompson, Deeks & Altman, 2003).

The results of both *Q* and I^2 tests, which are produced by the CMA software, were considered to decide whether an additional moderator analysis is needed. If effect size values were found to be heterogeneous (a large *Q* and I^2), a further examination of the nature of this variation and exploration of potential factors that might explain this variation is warranted by using moderator analysis.

Moderators Analysis

The third question of the present study was whether the following variables might moderate the relationship between TOEFL scores and GPA: publication status (published vs. unpublished); test version (pBT, cBT, and iBT); degree level (graduate or

undergraduate); and study setting (international or U.S.). To answer this question, a moderator analysis was conducted to determine what variables account for the differences among the effect size values.

To conduct the moderator analysis, the effect size values were first grouped according to the moderator of interest. For example, effect size values were classified into two main categories based on the publication status of the studies from which they were drawn. The moderator analysis was then conducted by the CMA software using the analog to ANOVA procedure in which a separate *Q* statistic is calculated for each subgroup (Lipsey & Wilson, 2001).

Each individual *Q* represents a weighted sum of squares of the studies in a certain subgroup about the mean of that subgroup. The sum of these individual *Q*s equal the within-group statistic: $Q_W = Q_1 + Q_2$. The within-group (Q_W) statistic is distributed as a chi-square with k − j degrees of freedom where k is the number of studies and j is the number of groups of studies.

The difference between the total *Q* total and the within-group *Q* is the between-group statistic: $Q_B = Q_T - Q_W$. The between-group (Q_B) statistic is also distributed as a chi-square with j−1 degrees where j is the number of groups of studies. If the between-group statistic for a certain moderator (e.g., publication status) is significant, it will be concluded that this variable accounts for the difference in effect size values. The smaller the (Q_W) value and the larger the (Q_B) value, the more the variability would be due to the moderators rather than error (Lipsey & Wilson, 2001).

Publication Bias

Although it is a powerful research method, meta-analysis is not without limitations (Thompson & Pocock, 1991). One main issue that impacts meta-analysis results is publication bias (Duval & Tweedie, 2000). This type of bias occurs when the sample of studies included in the meta-analysis is unrepresentative of the available studies of interests. The presence of publication bias in meta-analysis is due to various factors. Some of these factors are under the control of the researcher such as the case when the literature search is not exhaustive of all resources or includes only published research. In response to this problem, the current study employed an inclusive search strategy of all available research including both published and unpublished studies.

A more common source of publication bias is due to the fact that the majority of published studies are the ones that report statistically significant or positive results rather than the studies with non- significant or negative results known as the "file-drawer problem" (Rosenthal, 1979). This type of bias is out of the control of researchers since it is usually hard to access those unpublished studies. Even when these studies are found, their results are often reported as non-significant without providing the actual results that are needed for the meta-analysis. For example, Woodrow (2006) concluded that there was no significant relationship between TOEFL scores and GPA without reporting the actual correlations.

As a serious threat to the validity of meta-analysis, publication bias, regardless of its sources, should be addressed in analyzing the data. Publication bias might overestimate the effect size since most of the available studies are with significant

results. The following are the procedures that were employed to detect and evaluate the effect of publication bias in the present study.

Identification of Publication Bias

An effective graphical tool that used to visually detect the presence of publication bias was the funnel plot (Duval & Tweedie, 2000). This plot, which is generated by the CMA software, outlines each study effect size in relation to its measure of variability (e.g., sample size or variance). Publication would be unbiased when studies with large sample sizes have less variability (more precision) and studies with small sample size have more variability (less precision).

Publication bias would be presented if the funnel plot has a symmetric funnel shaped where studies with small sample sizes are not scattered around the average effect size but centered above the average. An asymmetric plot would indicate that the analysis is biased against non-significant results (Sterne & Egger, 2001). However, the use of these plots, though valuable, does not confirm the presence of publication bias (Egger, Smith, Schneider & Minder, 1997) since it is an exploratory tool for data inspection.

Statistical Procedures for Publication Bias

To determine the effect of publication bias on the results of meta-analysis, commonly statistical procedures were performed using the CMA software including the "Fail-safe *N*" test and "the Trim and Fill" method. The "Fail-safe *N*" test is used to estimate how many more studies would be added to the analysis to change the statistical significant results to non-significant (Rosenthal, 1979). Publication bias would be evident if the Fail-safe *N* test indicates that a small number of studies would be needed.

Since the Fail-safe *N* method has several drawbacks (Becker, 2005), a more powerful procedure used in this study is "the Trim and Fill" method (Duval & Tweedie, 2000). This method has the ability to detect the degree to which publication bias impacts the meta-analysis outcomes and how the effect size estimate would be changed after removing this bias. The Trim and Fill method is an iterative process in which studies are trimmed and filled and new estimates of the numbers of missing studies are computed. Adjustments of the effect size are made until an unbiased value of the effect size is estimated.

The results of visual and statistical methods of publication bias need to be interpreted with caution. The funnel plots might be asymmetric and statistical methods might indicate a bias even with the absence of publication bias.

Summary

In this chapter, the researcher reviewed the methods that were employed to conduct the meta- analysis of the relationship between the Test of English as a Foreign Language (TOEFL) and academic performance of international students as measured by grade point average (GPA). Detailed descriptions of the steps that were undertaken for conducting a meta-analysis were presented including: the search strategy, sample selection and analysis. The results of the meta-analysis are presented in the next chapter.

CHAPTER IV

RESULTS

In this chapter, the researcher presents the results of the meta-analysis conducted to investigate the predictive validity of international students' TOEFL scores on GPA. The results are organized in terms of the steps undertaken to carry out the meta-analysis. First, a description of the studies included in the analysis is presented. The findings of the meta-analysis are then discussed. The remaining sections include the results for the heterogeneity test, moderator analysis, and publication bias. The actual meta-analysis was conducted using the meta-analysis software, Comprehensive Meta-Analysis (CMA; Borenstein et al., 2005). In addition, SPSS and excel were employed to conduct additional analyses (e.g., exploratory and descriptive).

Search Results

The search process described in the previous chapter generated 41 studies forming the sample included in the meta-analysis. From the 41 studies, 47 independent effect size values were generated. This number of effect size values was due to the fact that some studies generated more than one effect size. As shown in Table 3, while the majority of authors reported single effect size values in their studies, three authors reported multiple effect size values in their studies. For example, Arcuino (2013) reported three separate effect size values drawn from three independent samples of students taking the three different versions of the TOEFL test.

Table 3: *List of Included Studies*

Author(s)	Year	*N*	Status	Level	Version	Setting
Al-Ansari	2003	90	Published	Undergraduate	PBT	International
Arcuino 1	2013	399	Unpublished	Graduate	IBT	U.S.
Arcuino 2	2013	241	Unpublished	Graduate	CBT	U.S.
Arcuino 3	2013	118	Unpublished	Graduate	PBT	U.S.
Burmeist	2014	108	Published	Graduate	CPT	U.S.
Carty	2007	34	Published	Undergraduate	PBT	U.S.
Chang	2009	378	Unpublished	Undergraduate	CBT	U.S.
Cho 1	2012	1,850	Published	Graduate	IBT	U.S.
Cho 2	2012	744	Published	Undergraduate	IBT	U.S.
Dunn	2006	203	Unpublished	Graduate	PBT	U.S.
Elliott	2011	16	Unpublished	Undergraduate	CBT	U.S.
Fass-Holmes	2014	328	Published	Undergraduate	IBT	U.S.
Fournier	2013	255	Published	Undergraduate	PBT	International
Fu 1	2012	245	Unpublished	Undergraduate	IBT	U.S.
Fu 2	2012	157	Unpublished	Undergraduate	PBT	U.S.
Fu 3	2012	628	Unpublished	Graduate	IBT	U.S.
Fu 4	2012	734	Unpublished	Graduate	PBT	U.S.
Gong	2006	165	Published	Undergraduate	PBT	U.S.
Hoefer	2000	590	Published	Graduate	PBT	U.S.
Itaya	2008	144	Published	Graduate	PBT	U.S.
Koys	2009	476	Published	Graduate	PBT	International
Kwai	2010	454	Unpublished	Undergraduate	PBT	U.S.
Lee	2005	88	Unpublished	Graduate	PBT	U.S.
Lo	2002	82	Unpublished	Undergraduate	PBT	U.S.
Maleki	2007	48	Published	Undergraduate	PBT	International
Melnick	2011	84	Published	Graduate	PBT	U.S.
Nelson	2004	844	Published	Graduate	PBT	U.S.
Ng	2007	433	Unpublished	Undergraduate	IBT	U.S.
Person	2002	72	Unpublished	Graduate	PBT	U.S.
Pitigoi-Aron	2011	137	Published	Graduate	PBT	U.S.
Poyrazli	2001	79	Published	Graduate	PBT	U.S.
Sahragard	2011	151	Published	Undergraduate	PBT	International
Sailor	2013	370	Unpublished	Undergraduate	PBT	U.S.
Salinas	2007	34	Unpublished	Graduate	PBT	U.S.
Seaver	2012	41	Unpublished	Undergraduate	IBT	U.S.
Simner	2007	345	Published	Undergraduate	PBT	International
Stacey	2005	171	Published	Graduate	PBT	U.S.
Takagi	2011	165	Unpublished	Undergraduate	IBT	International
Theuri	2007	54	Published	Undergraduate	PBT	International
Viravaidya	2007	157	Unpublished	Graduate	IBT	International
Vu	2011	464	Unpublished	Graduate	IBT	U.S.
Vu & Vu	2013	464	Published	Graduate	IBT	U.S.
Wait	2009	2,787	Published	Undergraduate	PBT	International
Wang	2013	575	Unpublished	Graduate	CBT	U.S.
Ward	2014	1341	Unpublished	Undergraduate	IBT	U.S.
Woodrow	2006	10	Published	Graduate	CBT	International
Zhang	2012	142	Unpublished	Graduate	IBT	U.S.

The included studies are listed alphabetically by the main author's last name in Table 3. The table also includes descriptive information for the publication year and sample size. The full codes for moderators are also listed: (a) publication status, (b) degree level, (c) test version, and (d) research settings.

Across the 41 studies, the total sample size was 17,495 participants. With an average of 372 participants, the sample size for studies included in the meta-analysis ranged from a minimum of 10 to a maximum of 2,787. As indicated by the median, almost 50 % of included studies have a sample size above 171. Descriptive data about the sample size for the studies in this meta-analysis are presented in Table 4.

Table 4

Frequency Distribution of Included Studies by Sample Size

Sample size	Frequency	Percent (%)	Cum. Percent (%)
30 or fewer	2	4.3	4.3
31-99	11	23.4	27.7
100-299	15	31.9	59.6
300-599	12	25.5	85.1
600-999	4	8.5	93.6
1000 or more	3	6.4	100.0
Total	47	100.0	

Publication Trends

As shown in Figure 3, most studies were published after 2007. This recent publication trend reflects the increased interest in the issue of TOEFL predictive validity. It is noteworthy that this number is not inclusive as the researcher located, through the literature review, about 68 studies that were published after 2000. As stated in the previous chapter, these studies were not included because they did not meet the criteria to be included in the meta-analysis.

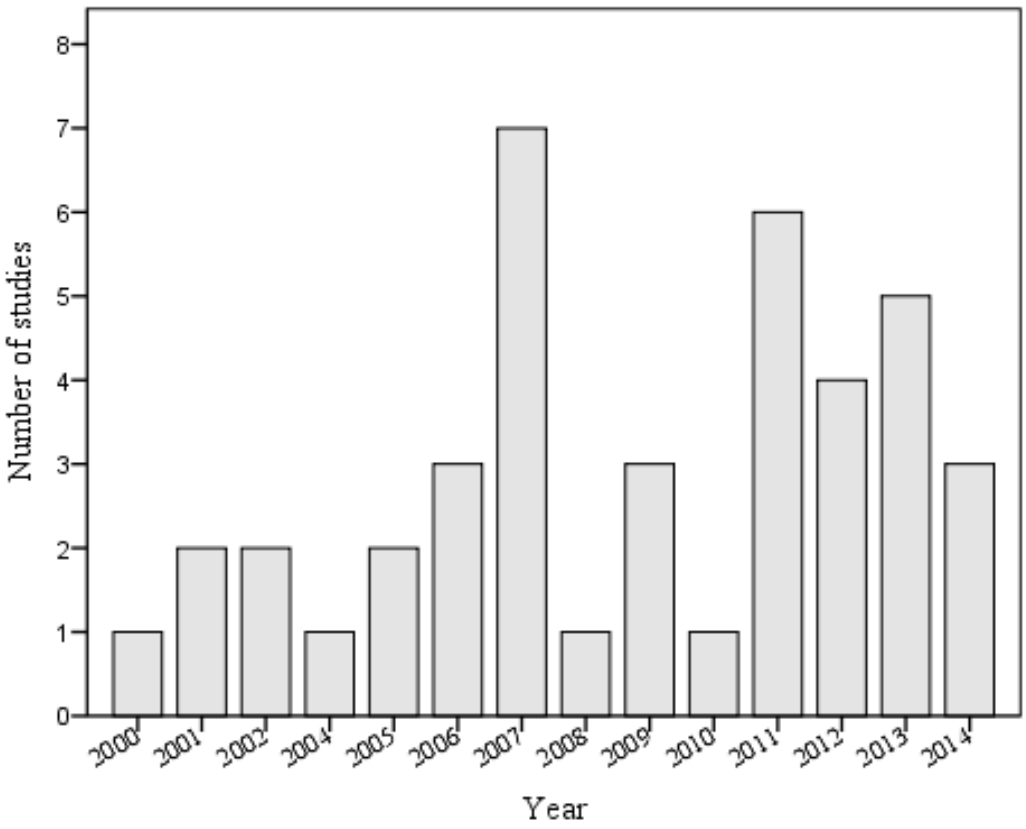

Figure 3. Number of Studies by Year of Publication

For the sake of showing the growing interest in the TOEFL predictive validity, a comparison of publication trend with the 1990s period was made. Figure 4 highlights the number of studies that were published in the 1990s period in comparison to the 2000 period. In comparison to the 68 studies that published after 2000, 24 studies that were published in 1990s. As exhibited in Figure 4, the number of studies peaked in after 2000.

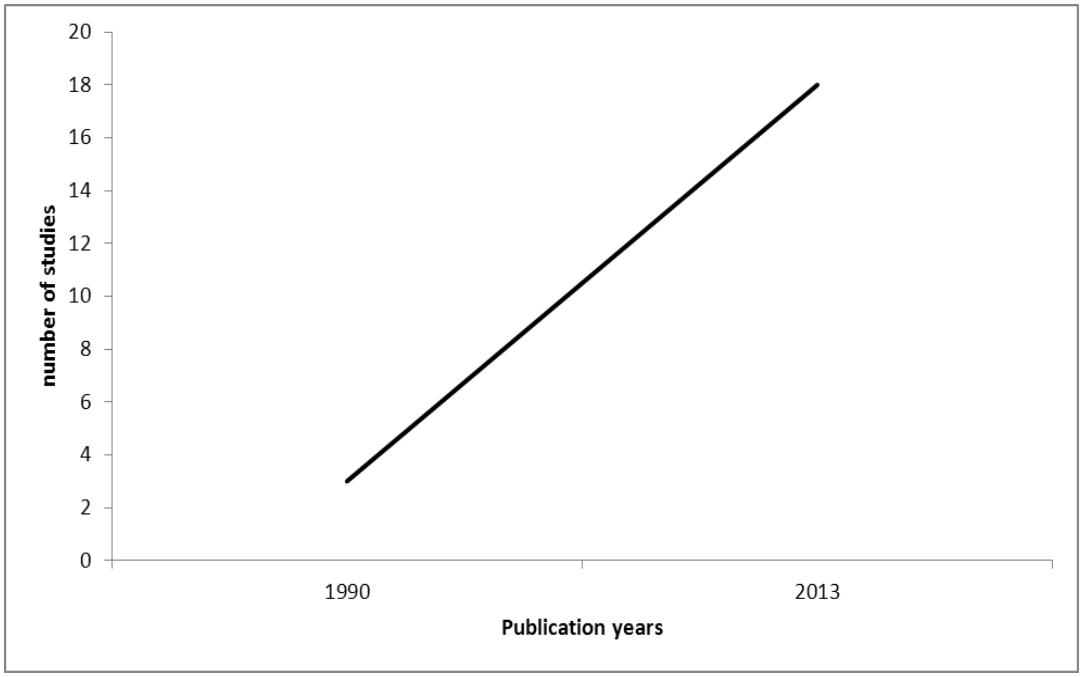

Figure 4. Number of Studies from 1990 – 2013

Effect Size

The Pearson product-moment correlation coefficient (referred to as Pearson's *r*) is the effect size index that was used in the analysis. Most of these effect size values (*r*) were obtained directly from studies ($n = 41$). The remaining effect size values ($n = 6$) were not reported in the studies and, therefore, were estimated using data available in those studies. The authors of these studies were contacted to get any results or data but it was not possible to get the direct correlations.

Four studies included regression coefficients in which the TOEFL was one of several predictors for the GPA. The regression coefficients express the correlation between TOEFL and GPA only in relation to other predictors included in the model (Lipsey & Wilson, 2001). In other words, if the regression model in which the correlation estimated was modified by adding or removing variables, the regression coefficient may be changed. Therefore, synthesizing these regressions as effect size in meta-analysis might yield inaccurate estimates (Kock & Gemünden , 2009). For example, Change and Agronow (2009) reported the regression coefficients for the effect of several admission predictors on the GPA of international undergraduate students. The regression model included multiple predictors (TOEFL, SAT, high school GPA, and gender). The authors found a regression coefficient of (.0069) for the TOEFL effect on GPA.

There is a debate on synthesizing the regression coefficients in meta-analysis of correlations; some researchers argue against using the regression coefficient (e.g, Hunter and Schmidt, 2004) while others call for including them in the meta-analysis of

correlations (e.g, Peterson & Brown, 2005). The reader should be aware that the purpose of this analysis was not to argue for or against any of these methods but to find the method that yields the most accurate results.

As the researcher of the current meta-analysis sought to include the maximum number of available studies, the researcher decided to include the regression coefficients in the analysis by converting them to (*r*) using the following imputation suggested by Peterson and Brown (2005).

$$r = .98\beta + .05\lambda$$

Where β is the regression coefficient of interest, and λ is a parameter equals 1 when the coefficient is positive and zero when it is negative. An assumption of this rule is that the value of the β should be between -.50 and .50. This assumption was met for the four regression coefficients included in this meta- analysis. For example, the regression coefficient in Change & Agronow (2009) study described above was converted to a correlation coefficient by:

$$r = .98*.0069 + .05*1 = 0.057$$

According to Kock and Gemünden (2009), researchers should be careful in applying this imputation because of the error that might result from this approximation. To avoid such error, the meta-analysis was conducted with and without these regression coefficients, and then their estimates were compared in order to assess their impact on the average effect size (R. A. Peterson, personal communication, November 06, 2013).

Another issue with synthesizing effect size values occurs when different studies use the same data. In the current study, only the two studies by Vu (2011) and Vu &Vue

(2013) used the same data and reported the same results. The only difference between these two studies is the year and type of publication, Vu (2011) is an unpublished dissertation and Vu &Vue (2013) is published article. Both of the studies were included in the meta-analysis since they have different publication status which will be useful in conducting the moderator analysis. As with the case of regression coefficients described above, the meta-analysis was conducted with and without the latter study and estimates of the two average effect size values were compared.

One of the challenges with meta-analysis is that the results of some studies are reported in terms of statistical significance without providing the actual correlations or other statistical tests that allow the estimation of effect size from the reported data. This meta-analysis included one study by Woodrow (2006) that reported its results about the correlation between the TOEFL and GPA as a non-significant with providing any additional data. As a recommended practice in meta-analysis, Woodrow's (2006) study was included in the analysis with an estimated effect size of $r = .0$ corresponding the lowest possible effect size (Card, 2013).

Exploratory Analysis

The results of meta-analysis are valid as far as the data on which the analysis is based, namely the individual effect size values. Therefore, examination of the nature and the distribution for the effect size values occurred prior to conducting the meta-analysis. The distribution of correlations is usually positively skewed, thus violating the normality assumption required to synthesize effect size values in meta-analysis (Rosenthal, 1991).

To normalize this distribution, Fisher's *r* to *z* transformation was applied. The distribution of transformed *r to z* values is displayed in Figure 5. The difference between *r* and z becomes noticeable in the higher values. For example, an *r* of .140 has a *z* value of .141 whereas an *r* of .6 has a *z* value of .7. The analysis was conducted using these transformed *z* effect size values.

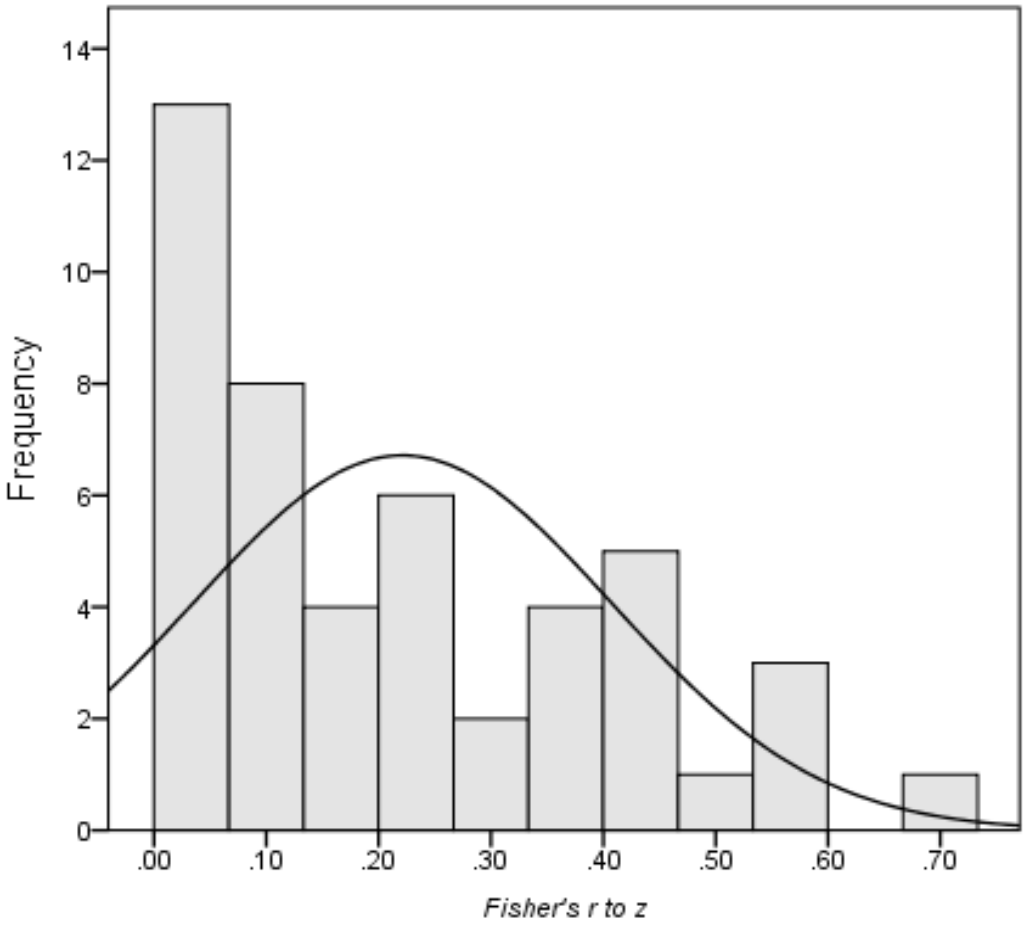

Figure 5. Distribution of Fisher's *r to z* Values Reported From Studies

A visual description of the effect size values variability and distribution is shown in the stem-and-leaf in Table 5. As presented in this table, the majority of the effect size values (n= 32) is distributed between the values of 0.0 and 0.3.

Table 5

Stem-and-leaf Plot for Effect Size Values

Stem	Leaf	Count
0.0	0 0 1 1 2 2 4 4 5 5 5 6 6 7	14
0.1	0 1 2 2 2 2 3 4 4 6 8	11
0.2	0 2 2 3 5 5 7	7
0.3	0 4 6 7 8	5
0.4	1 2 2 4 5	5
0.5	2 5 8 9	4
0.6		0
0.7	1	1

More information about the distribution of effect size values was revealed through checking the boxplot presented in Figure 6. The lowest extreme value is .0 and the highest effect size is .71. The median is .16 indicating that 50% of the effect size values are above .16, 75% are below .37, and 25% are below .06. In addition, quartiles were used to detect outliers in the data. The researcher, through a review of the effect size values, found no outliers in the data.

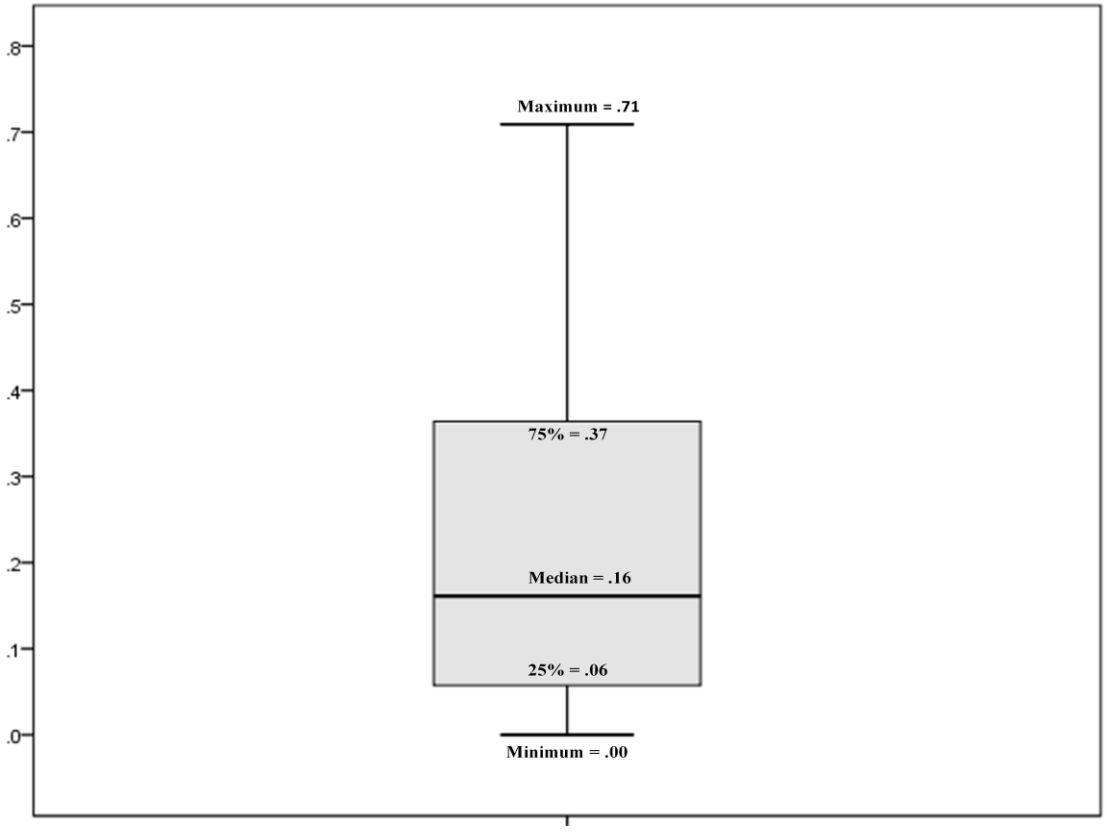

Figure 6. Box Plot of Effect Size Values Reported From Studies

Meta-Analysis Findings

As described previously, the following procedures were followed to conduct the meta-analysis. Individual effect size values were directly obtained or estimated from each included study. Then, the overall summary effect size was computed using both fixed and random model. In the next stage, the variability of effect size values across studies was examined using the heterogeneity test statistics. A follow up analysis of potential moderator variables was also conducted.

The current meta-analysis was conducted following the procedures described by Hedges and Olkin (1985). The random effect model, as described in the previous chapter, was chosen as the model for analyzing and interpreting the effect size values. However, the results for both of the fixed and random effect model are presented to allow a comparison of the two models.

To answer the first question concerning the relationship between TOEFL scores and academic performance as measured by GPA, the weighted average of the 47 effect size values was computed. As summarized in Table 6, the findings of the meta-analysis yielded an overall effect size ($\overline{ES}$) of .21. This effect size indicates that across all effect size values drawn from studies, there is a positive correlation between the predictor (TOEFL) and the outcome (GPA). Following the criteria suggested by Cohen (1988) for interpreting the effect size (*r*), an effect size of .21 falls within a small to medium range.

Table 6

Results from Meta-Analysis

					95% CI	
Model	*N*	*k*	$\overline{ES}$	Se	Lower	Upper
Random	17,495	47	0.21	0.02	0.16	0.26
Fixed	17,495	47	0.16	0.01	0.14	0.18

Note: *N* = total sample size; *k* = number of effect size values; $\overline{ES}$ = overall weighted effect size; Se = standard error; CI = 95% upper and lower confidence intervals around the effect size.

Confidence Intervals

Using the standard error (Se = 0.025), confidence intervals were obtained to estimate the extent to which such intervals would include the population average effect size.

Lower 95% CI = .21 - (1.96 × .025) = .16

Upper 95% CI = .21 + (1.96 × .025) = .26

The obtained CI values indicate that 95% /100 such interval (.16 and .26) would include the population average effect size. This confidence interval is narrow and a precise estimate of the effect size. As the confidence interval (.16 - .26) does not include zero, it is concluded the average effect size $\overline{ES}$ = .21 is statistically significant indicating there is a relationship between the TOEFL and GPA.

The interpretation of effect size values can also be enhanced by examining the position and width of their corresponding confidence intervals. The 47 effect size values and confidence intervals for individual studies as well as the total average effect size are

summarized in the forest plot in Figure 7. In this plot, studies are listed along with their effect size values which are represented by boxes and the widths of confidence intervals are highlighted by lines. As displayed in the plot, the different sizes of the boxes are a visual reflection of the weights of each individual study in the overall effect size; a large box indicates the study has large contribution to the overall effect size estimate. The overall average effect size (for the random-effect model) is summarized at the bottom of the plot and represented by a diamond.

Comparing the two models, the random-effect model generated a larger effect size ($\overline{ES}$ = .21) than the fixed effect model ($\overline{ES}$ = .16). Also, the random-effect model produced larger a confidence interval due to the addition of a variance component (i.e., the between study variance). The difference between these two models might be due to the fact that each model assigns a different weight for individual studies. Under the fixed effect model, studies with larger sample sizes have more contribution to the average effect size than small studies. For example, the study by Person (2002) with a sample size of 72 has a weight of only 0.47% under the fixed model while its weight under the random model is 2%. On the other hand, the study by Wait (2009) with a larger sample size of 2,787 has a weight of 19% in the fixed model and a weigh of 3% under the random models.

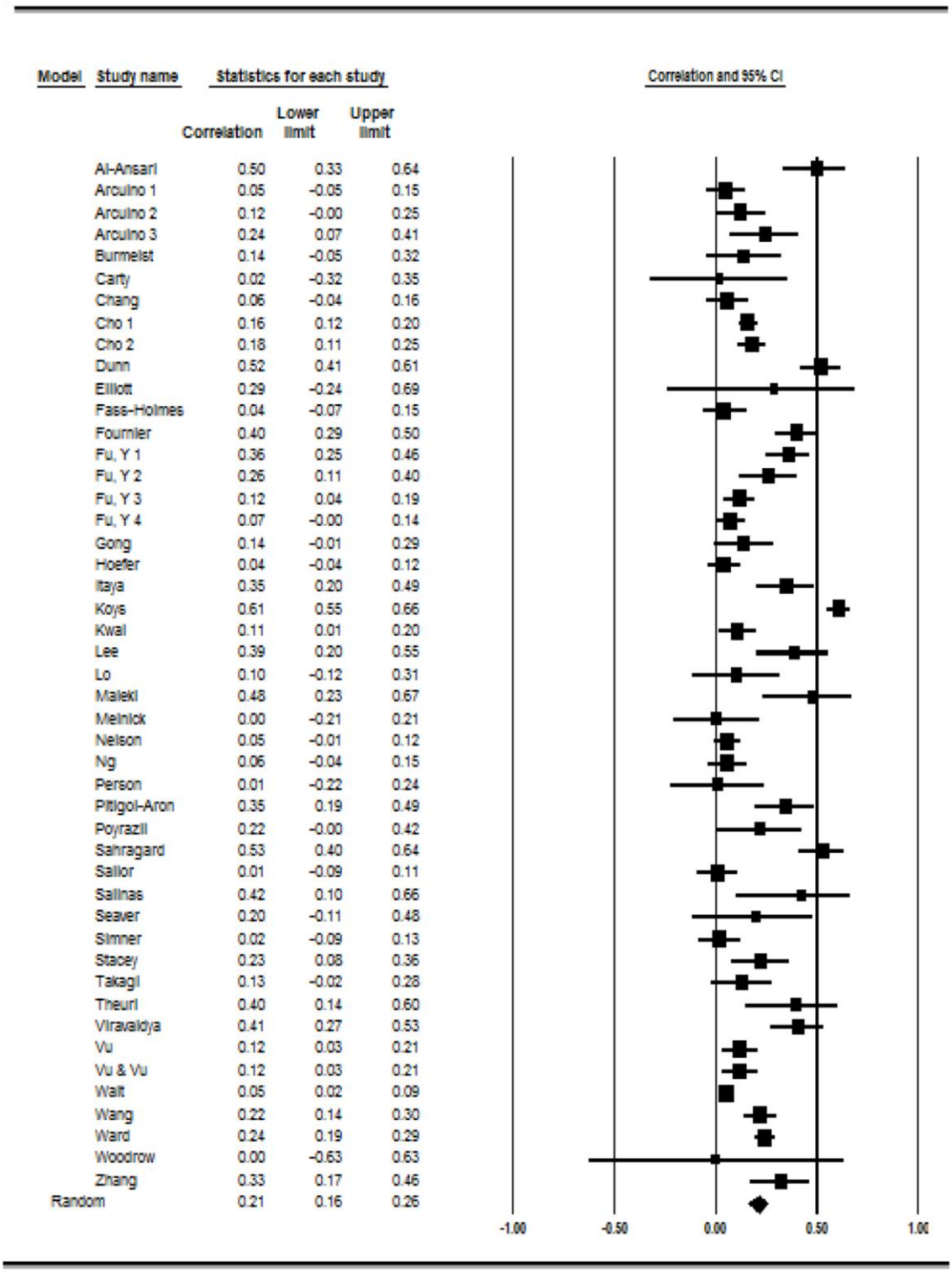

Figure 7. Forest Plot of Effect Size Values Reported From Studies in Meta-Analysis

Confidence intervals can also give insights about the contribution of each effect size in the overall effect size. The wider the confidence intervals, the less precise and less weight the effect size has. This fact is reflected in the forest plot in Figure 7; the effect size values with wide CIs have smaller boxes. An important fact that is revealed from examining this plot is that the confidence intervals of approximately 26 (55%) studies do not cross the zero indicating that these studies have statistically significant effect size values while the remaining 21 (45%) studies that cross the zero have a non-significant effect size values.

As was previously discussed, the analysis was repeated excluding some studies that might have some issues in estimating the effect size. The studies excluded from this analysis include those that reported regression coefficients (n= 4) as well as studies with similar data (n = 1). In comparison with the average effect size estimate of .21, the results of this repeated analysis yielded a slightly higher effect size estimate of $\overline{ES}$ = .24.

Heterogeneity of Effect Size

To answer the second question; whether effect size value are homogenous or heterogeneous across studies, graphical and statistical procedures were employed to investigate variation of effect size values. The forest plot in Figure 7 was first examined to find about the variability of effect size values. The confidence intervals for about 23 studies overlap with the average effect size demonstrating that these effect size values are similar in their true effect size. However, the confidence intervals of the remaining 24 studies do not overlap with the average effect size showing a large amount of variability in these effect size values.

The *Q* test (Hedges & Olkin, 1985) was then conducted. The *Q* statistic estimates the weighted sum of squared deviations of each effect size from the weighted average effect size. The larger the *Q* value, the more variable the effect size values, and the wider the confidence intervals around the effect size values.

As summarized in Table 7, the *Q* statistic is significant (χ^2 (46) = 406, $p < 0.001$) and evaluated using the corresponding critical value from the chi-square distribution with (47-1) degrees of freedom. As the *Q* statistic exceeds this critical value, the null hypothesis that the effect size values are homogenous is rejected and it is concluded that the variation between effect size values is larger than what would be expected by chance (sampling error) alone.

Table 7

Results of Heterogeneity Test

Heterogeneity test			Variance		
Q	*df*	*p*	I^2	95% CI for I^2	
				Lower	Upper
406	46	0.000	89%	86%	91%

Note: Q = the obtained Chi square value; *df* = the degree of freedom which equals the number of effects sizes minus 1(47-1); *p* = significant level; I^2 = the amount of the true variance between effect size values; CI = 95 % upper and lower confidence intervals around I^2 index.

To estimate the amount of the true variance between effect size values (i.e., the degree to which these effect size values vary), the index I^2 was computed as the ratio of the between study variance to the total variance. Large I^2 indicates heterogeneity and its value will be interpreted according to the following criteria suggested by Higgins, Thompson, Deeks & Altman (2003): I^2 = 75%: large heterogeneity; 50%: moderate heterogeneity; and 25%: low heterogeneity. As presented in Table 7, The I^2 is 89% which represnets a large amount of heterogeneity for the effect size values.

As the results of both Q and I^2 tests indicate substantial heterogeneity in effect size values, a further examination of the nature of this variation and exploration of potential factors that might explain this variation is warranted using moderator analysis.

Moderators Analysis

The third question of the present study examined whether the following variables might moderate the relationship between the TOEFL scores and GPA: publication status (published vs. unpublished); test version (pBT, cBT, or iBT); degree level (graduate or undergraduate); and study setting (international or U.S.). To answer this question, a moderator analysis was conducted to determine whether and to what degree these variables account for the differences among the effect size values.

To conduct the moderator analysis, the effect size values were grouped according to the moderator of interest. For example, effect size values were classified into two main categories based on the publication status of the studies from which they were drawn. The analysis was conducted on these subgroups using the analog to ANOVA

procedure in which a separate *Q* statistic is calculated for each subgroup (Lipsey & Wilson, 2001).

The results of the moderator analysis are presented in Table 8. Summary statistics of each subgroup are first examined. There were 23 (49%) published studies and 24 (51%) unpublished studies. Of these unpublished studies, the majority (n = 17) were dissertations, four master theses, one conference presentation, and two research reports. Figure 8 highlights the number of studies per year by publication status. As shown, recent studies are mostly unpublished indicating a growing number of dissertations conducted on the topic.

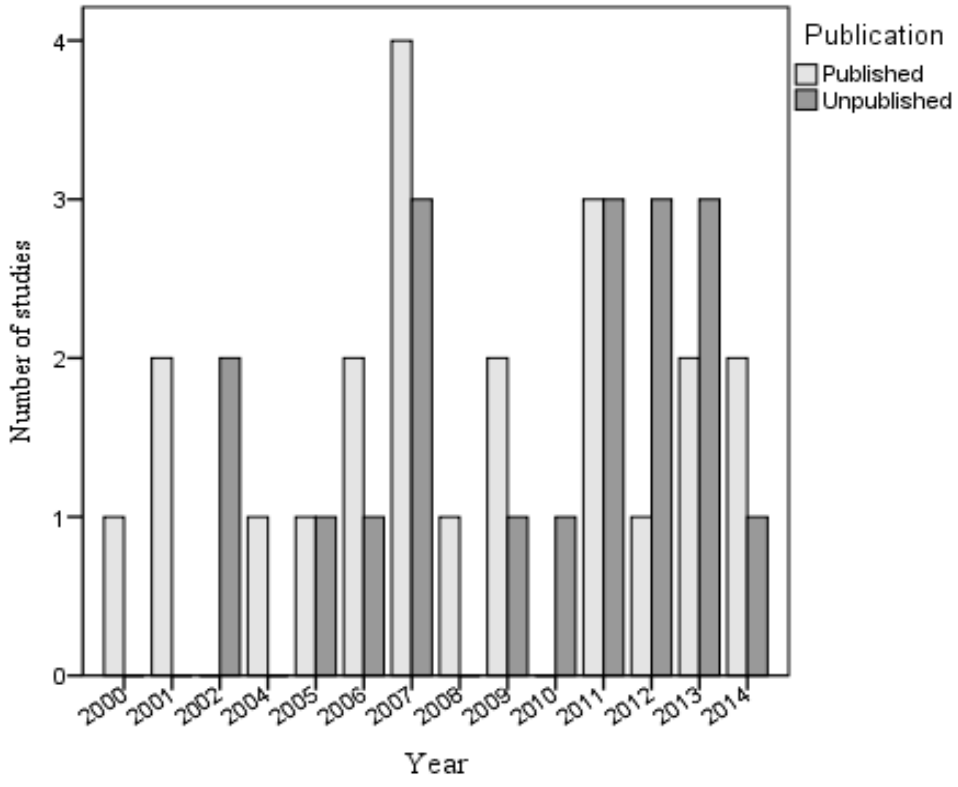

Figure 8. Publication Status per Year for Studies in the Meta-Analysis

As shown in the summary statistics in Table 8, while the TOEFL pBT test is the most frequent version used in studies, few studies conducted on the TOEFL cBT version. These figures are due to the fact that pBT version was the longest test version in use, the cBT version used for a short period between 1998-2006 and is no longer in use; and the iBT is the most recent version.

The *Q*-test was examined then to find out whether the variation is due to a real difference between studies. As displayed in Table 8, the between groups *Q* statistic was not significant for three of the moderators: publication status ($\chi 2$ (1) = 0.7, $p > 0.05$) graduate level ($\chi 2$ (1) = 0.2, $p > 0.05$); and test version ($\chi 2$ (2) = 4.9, $p > 0.05$). The Q-test was evaluated using the corresponding critical value from the chi-square distribution with (*df*-1) degrees of freedom. Since the Q statistic did not exceed this critical value, the null hypothesis that the effect size values are homogenous was failed to be rejected and, therefore, there is no evidence that the effect size values are different for these three moderators. In other words, the difference in the effect size values between moderators is due only to chance rather than to a true difference.

The heterogeneity test for the research setting indicated that the between groups *Q* statistic was significant ($\chi 2$ (1) = 3.9 $p < 0.05$) and evaluated using the corresponding critical value from the chi-square distribution with (2-1) degrees of freedom. Since the *Q* statistic exceeds this critical value, the null hypothesis that the effect size values are homogenous was rejected and, therefore, there is evidence that the difference between the effect size values of studies conducted inside and outside the U.S. is statistically significant.

Table 8

Results of Moderator Analysis

	Summary Statistics					Heterogeneity Test		
Moderator	*N*	*k*	$\overline{ES}$	Se	95% CI	*Q*	*df*	*p*
Publication Status						0.7	1	0.407
Published	9,985	23	0.23	.04	0.15 - 0.31			
Unpublished	7,537	24	0.19	.03	0.14 - 0.24			
Degree Level						0.2	1	0.639
Graduate	8,812	25	0.22	.04	0.15 - 0.29			
Undergraduate	8,683	22	0.20	.03	0.13 - 0.26			
Test Version						4.9	2	0.087
pBT	8,396	26	0.26	.04	0.17 - 0.34			
iBT	7,771	15	0.16	.03	0.11 - 0.21			
cBT	1,382	6	0.14	.04	0.07 - 0.21			
Research Setting						3.9	1	0.048
International	4,538	11	0.35	.09	0.17 - 0.51			
U.S.	12,975	36	0.17	.02	0.13 - 0.20			

Note: N = total sample size; *k* = number of effect size values; $\overline{ES}$ = overall weighted effect size; Se = standard error; 95% CI = upper and lower confidence intervals around the effect size; *Q* = the obtained Chi square value; *df* = the degree of freedom which equals the number of groups minus 1; *p*= significant level.

The results above indicate that the research setting is a potential moderator of the relationship between the TOEFL and GPA. A follow up analysis was conducted on this moderator to find out more about the nature and the factors that might lead to this variation. More specifically, the analysis focuses on the interaction of the research setting with other factors as presented in the following section.

The Interaction of Research Setting with Sample Size

Descriptive data about the percentages of sample size per setting are shown in Table 9. The average sample size for the international studies group is 413 and ranged from 10 to 2787. The median is 157 indicating that almost 50 % of the included studies have a sample size above 157. For the U.S studies group, the average sample size is 360 and ranged from 16 to 1850. The median is 222 indicating that almost 50 % of the included studies have a sample size above 222. These data reveal that U.S. studies had relatively larger sample sizes than international studies.

Table 9

Frequency Distribution of Included Studies for Sample Size by Setting

	International setting			U.S. setting		
Sample size	Frequency	Percent (%)	Cum. Percent (%)	Frequency	Percent (%)	Cum. Percent (%)
30 or fewer	1	9.1	9.1	1	2.8	2.8
31-99	3	27.3	36.4	8	22.2	25.0
100-299	4	36.4	72.7	11	30.6	55.6
300-599	2	18.2	90.9	10	27.8	83.3
600-999	0	0.0	90.9	4	11.1	94.4
1000 or more	1	9.1	100.0	2	5.6	100.0
Total	11	100.0		36	100.0	

The Interaction of Research Setting with Publication Year

Another analysis was conducted to determine whether there was an interaction between publication year and research setting. As shown in Figure 9, there are more recent U.S. studies than international studies. This could be due to the fact that U.S. studies become available sooner than international studies.

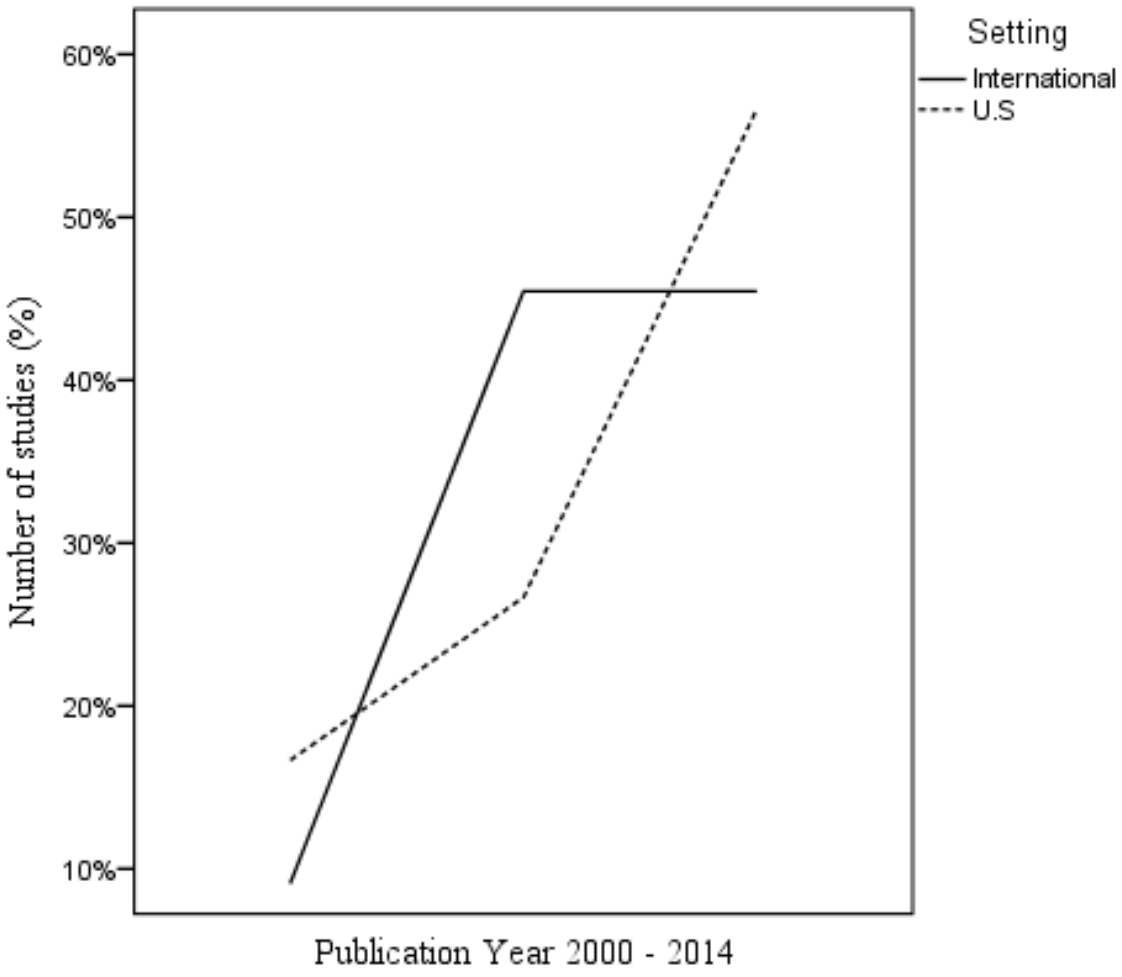

Figure 9. Interaction of Setting with Year

The Interaction of Research Setting with Publication Status

As presented previously, the number of studies conducted in the U.S. is considerably larger than the number of studies conducted outside the U.S. The percentages of studies by publication status for each study setting are displayed in Figure 10.

The majority (82%) of international studies are published whereas the majority of U.S. studies are unpublished (57%). The small number of unpublished international studies might be due to the difficulty in locating such studies.

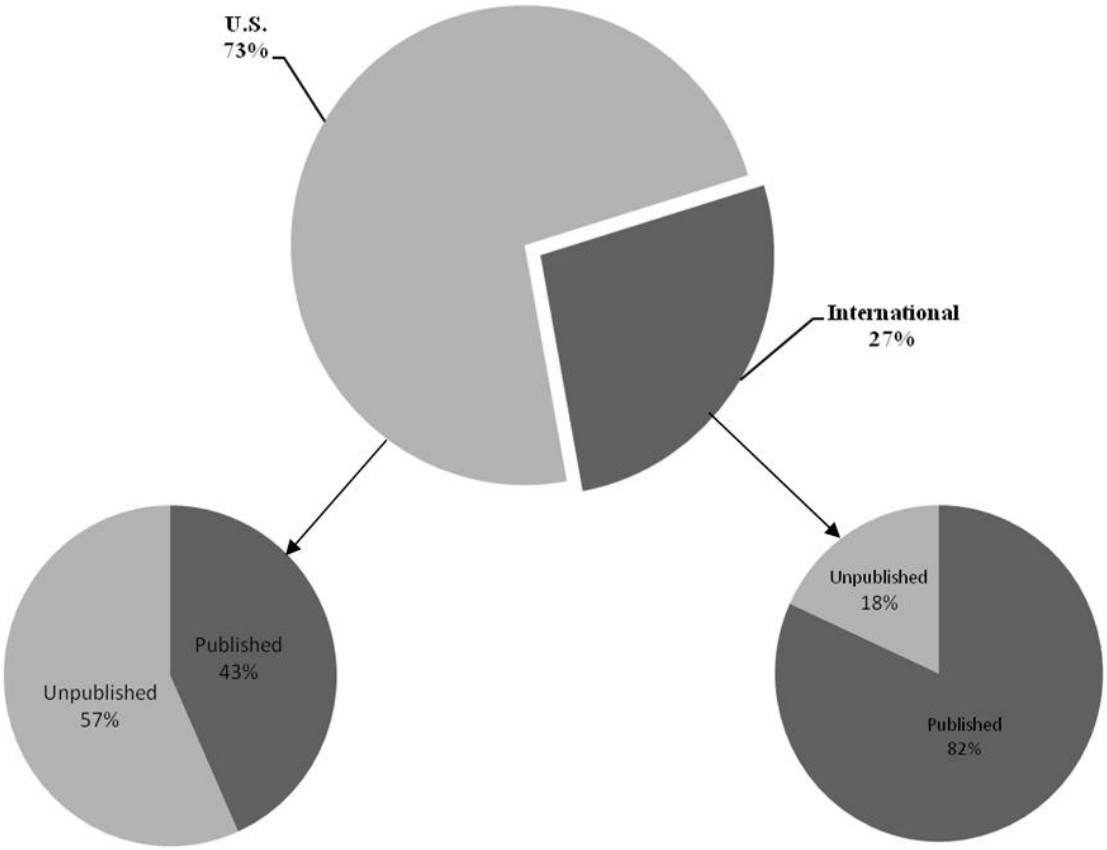

Figure 10. The Percentage of Studies by Publication Status for Each Setting

The researcher also examined the interaction between research setting and publication status. Put differently, does publication status (published or unpublished) vary based where the study was conducted (in or outside the U.S.)? Some insights about the nature of this interaction are illustrated in Figure 11. First, international studies have a higher effect size than U.S studies in both published and unpublished studies. Second, for the international setting, published studies have larger effect size than unpublished studies. The interaction is demonstrated in the fact that published studies have larger effect size in the international setting, whereas, unpublished studies have larger effect size in the U.S. setting.

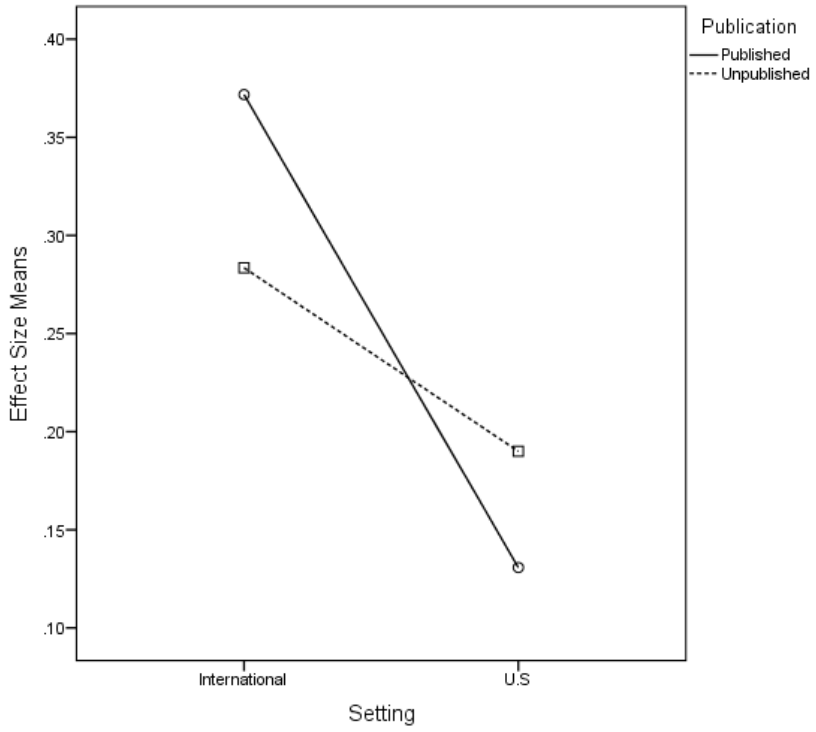

Figure 11. Interaction of Setting with Publication Status

The Interaction of Research Setting with Degree Level

Figure 12 highlights the interaction between setting and degree level. First, international studies have a higher average effect size than U.S studies for both graduate and undergraduate levels. Second, while graduate level has a larger average effect size than undergraduate level in the international setting, there is no or little difference between the two levels in the U.S setting.

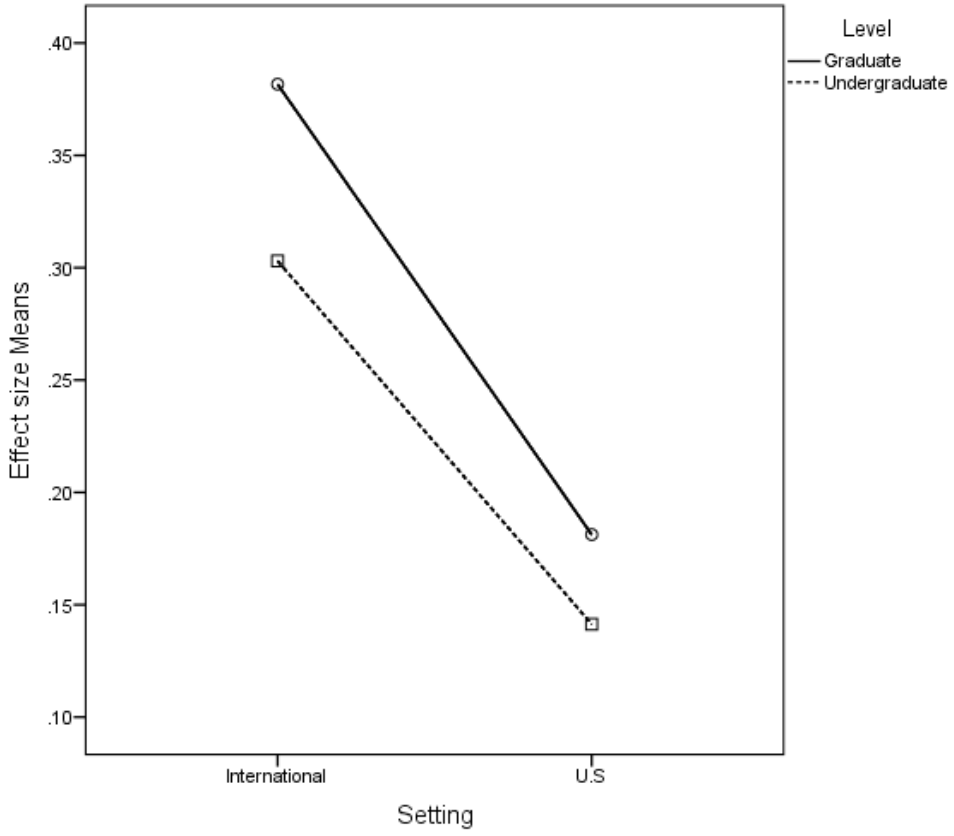

Figure 12. Interaction of Setting with Degree Level

The Interaction of Research Setting with Test Version

As displayed in Figure 13, international studies have higher average effect size than U.S studies across all test versions. Another fact that may be revealed by examining this figure is that the pBT test version has a larger average effect size than both cBT and iBT in the international setting. However, this variation between effect size values based on the test version is negligible in the U.S. setting. Moreover, there is no an interaction between study setting and test version.

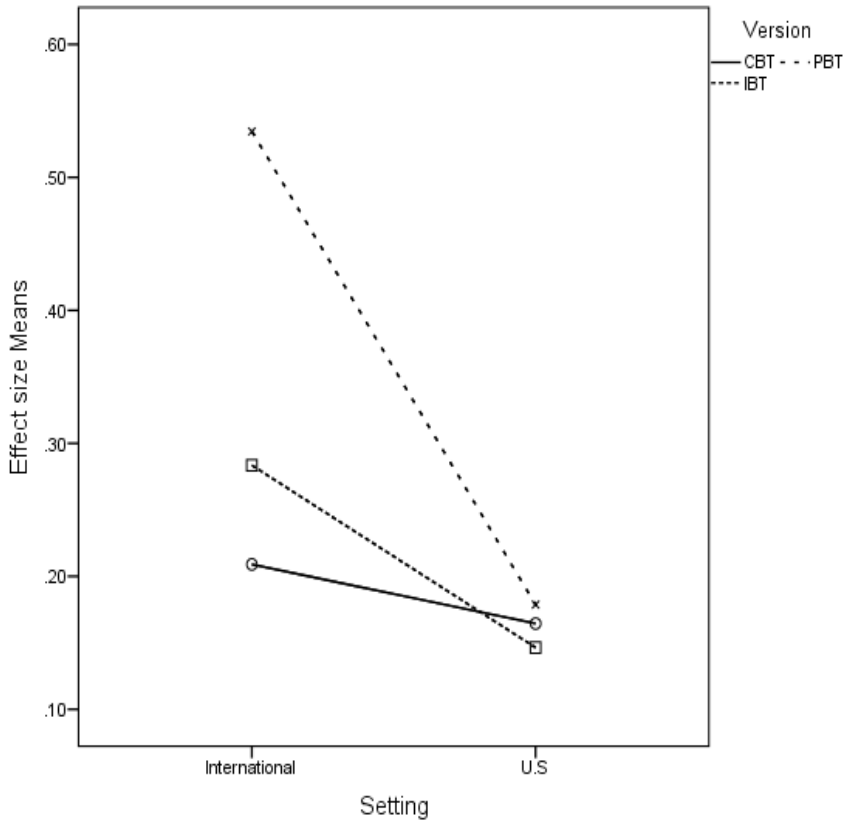

Figure 13. Interaction of Setting with Test Version

In summary, the interaction between study setting and the different moderators is apparent in the international setting rather than in the U.S one. However, due to the small number of international studies, these results therefore need to be interpreted with caution.

Publication Bias

As a serious threat to the validity of meta-analysis, publication bias, regardless of its source, should be addressed in data analyzing in meta-analysis. Since studies with significant results are more likely to be published than studies with non-significant results, publication bias might overestimate the effect size. In response to this problem, the researcher employed an inclusive search strategy of all available research including both published and unpublished studies. In addition, a variety of procedures were employed to detect and evaluate the effect of publication bias in the present study as discussed in the following section.

Identification of Publication Bias

One procedure that was utilized to visually identify the presence of publication bias is the funnel plot (Duval & Tweedie, 2000). The funnel plot in Figure 14 outlines each study effect size in relation to its standard error. The unfilled circles represent the observed effect size values while the filled circles represent the imputed studies generated from the trim and fill procedure. The unfilled diamond represents the observed average effect size while the unfilled diamond represents the adjusted average effect size resulted from the trim and fill procedure. The trim and fill procedure is discussed at the

end of this section. The mid line represents the mean effect size while the two solid lines represent the confidence interval.

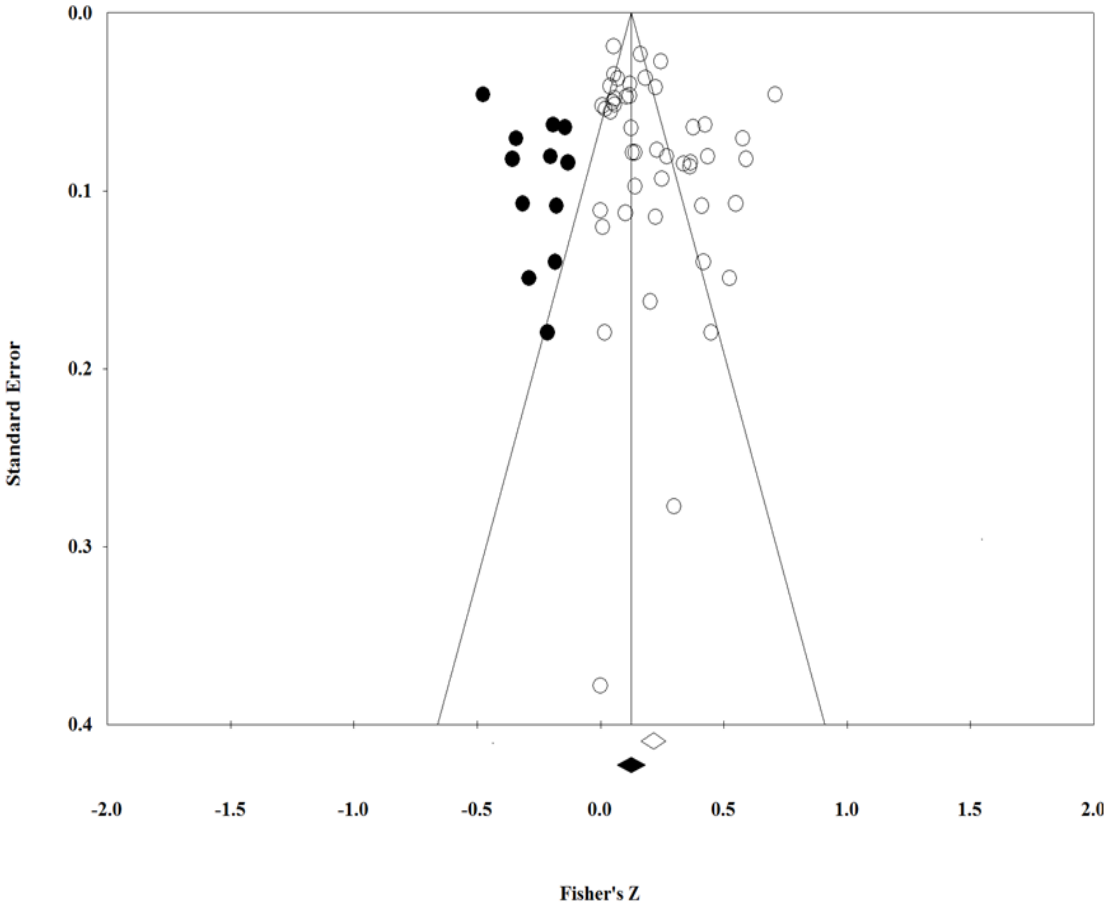

Figure 14. Funnel Plot of Effect Size Values by Standard Error

As displayed in Figure 10, studies with large sample sizes have small standard error and thus are centered at the top of the figure (more precision) and clustered around the mean effect size (less variability). On the other hand, studies with small sample sizes have large standard error (less precision) and thus are scattered around the mean effect size (more variability). The distribution of effect size values in this plot seems to be symmetric and therefore has less indication of publication bias.

Since the funnel plot above generated by the CMA software was not clear as to the detection of publication of bias, the researcher created a more accessible illustration that might help in assessing this bias and is presented in Figure 15. As highlighted in this figure, the distribution of individual effect size values (represented by dots) around the mean effect size (represented by dotted line) in relation to the study sample size (*N*). The lines divide the figure into 4 areas: the upper areas (A&B) include studies with large *N* while the lower areas (C&D) include the studies with small *N*. The left side areas (A&C) include studies with small effect size values (below the mean) while the right sides areas (C&D) include studies with large effect size values (above the mean). For example, the area C includes the group of studies that have small *N* and small effect size.

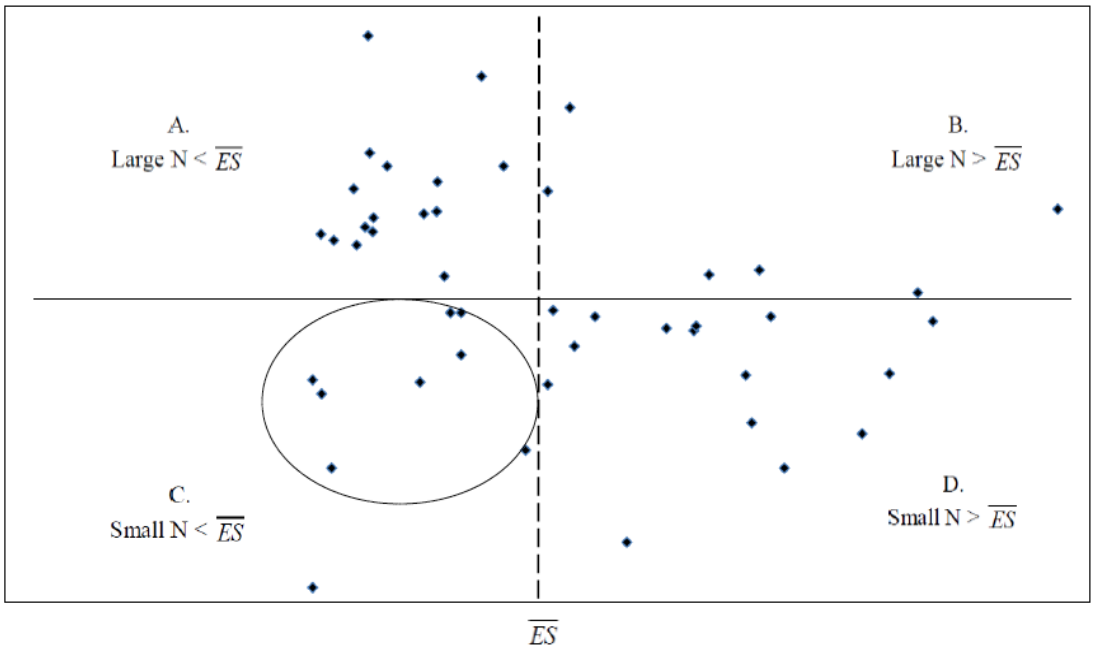

Figure 15. The Distribution of Effect Size (ES) Values in Relationship to Sample Size (*N*)

In order to assess the publication bias, we are interested with the number of studies with small *N* and small effect size (area C) that includes the studies with non-significant results. If there were no publication bias, we would expect enough number of these non-significant studies. As illustrated in this figure, the number of studies with non-significant results (highlighted by the circle in area C) suggests no strong indication of publication bias. Another interesting fact that was revealed from this figure is studies with smaller *N* tend to have a larger effect size.

The funnel plot above is an exploratory tool for data inspection. The interpretation of this plot is susceptible to subjective judgment and does not confirm the presence of publication bias (Egger, Smith, Schneider & Minder, 1997). Therefore, more rigorous statistical procedures for detecting Publication Bias are warranted.

Statistical Procedures for Detecting Publication Bias

To determine the effect of publication bias on the results of meta-analysis, a commonly used procedure is the "Fail-safe *N*" which estimates how many studies would be added to the analysis to change the statistical significant results to non-significant (Rosenthal, 1979). The results of Fail-safe *N* test generated by the software indicate that 5,704 more studies would be needed to change the effect size estimate. It is very unlikely that this number of studies would be missing and therefore the results are robust to publication bias.

Since the Fail-safe *N* method has several drawbacks (Becker, 2005), a more powerful procedure, the Trim and Fill, method was also used (Duval & Tweedie, 2000). This method has the ability to detect the degree to which publication bias impacts the meta-analysis outcomes and how the effect size estimate would be changed after removing this bias. A visual representation of "the Trim and Fill" test results was depicted in Figure 14. The observed studies are represented by the unfilled circles while the imputed studies generated from the trim and fill procedure are represented by the filled circles (*N*=12). Adjustments of the observed effect size (represented by the unfilled diamond) are made until an unbiased value of the effect size is estimated (represented by

the filled diamond). The statistical results of "the Trim and Fill" test generated by the software are presented Table 10.

As shown in Table 10, the "the Trim and Fill" method generated 12 missing studies. For the random-effect model, this method generated a new adjusted effect size estimate of .12 in comparison with the actual observed effect size of .21. While there is a difference between the observed and the adjusted effect size estimate, this difference might not be an indication of the presence of publication bias, especially if the confidence intervals of the two estimates are considered.

Table 10

Results for the "Trim and Fill" Test

		Fixed effects		Random effects	
Model	Studies Trimmed	$\overline{ES}$	*95 % CI*	$\overline{ES}$	*95 % CI*
Observed		.16	.14 - .18	.21	.16 - .26
Adjusted	12	.12	.10 - .13	.12	.07 - .18

$\overline{ES}$ = overall weighted effect size; CI = 95 % upper and lower confidence intervals around the effect size

Previous exploratory analysis also supports the notion of the absence of publication bias in this meta-analysis. For instance, 26 (55%) effect size values are below the mean whereas the remaining 21(44%) are above the mean. Additionally, there is almost equal number of unpublished (*n*= 24) and published studies (*n*=23). Assessing publication bias in relation to research setting revealed that publication bias could be more presented in international studies. It was found that seven of the largest effect size values (above the mean) are international studies. In summary, the results of the visual and statistical methods of publication bias presented above indicate no strong evidence for the presence of publication bias in this meta-analysis.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The contradictory results from previous research on the association between international students' TOEFL test scores and academic performance as measured by GPA highlights the importance of the current study's comprehensive examination of research regarding the TOEFL test and international students' academic performance. Adopting current meta-analytic methods, the researcher attempted to combine and compare the results of previous research on the relationship between TOEFL scores and GPA.

In this meta-analysis, the researcher both quantified the magnitude of effect size and examined different moderator variables that influence the relationship between students' TOEFL scores and GPA. Examining these variables could help to explain the variance in the observed relationships and give more insights on the contradictory results from previous research. These moderator variables include publication status, degree level, test version, and study setting.

The discussion in this chapter is organized by the following sections: (a) findings from the meta-analysis, (b) implications, (c) limitations, and (d) recommendations for future research.

Findings from the Meta-Analysis

The current study sought to answer the following research questions:

1. How valid are international students' TOEFL scores in predicting academic performance?
2. Do effect sizes vary across studies?
3. Which of the following factors moderate the TOEFL-GPA relationship: publication status (published vs. unpublished), degree level (graduate vs. undergraduate), test version (internet-based test (iBT) vs. computer-based test (cBT) vs. paper-based test (pBT), and research setting (U.S. vs. International)?

How Valid are International Students' TOEFL Scores in Predicting Academic Performance?

The first question was investigated by estimating the magnitude and direction of the effect size for all TOEFL-GPA correlations identified in the literature. Across the sample of 47 independent effect size values, an overall effect size of .21 (95% CI = .16 - .26) was identified. As the confidence interval does not include zero, it is concluded that this meta-analysis yielded a statistically significant and positive relationship between international students' TOEFL scores and their academic performance as measured by GPA.

According to the criteria suggested by Cohen (1988), an effect size of .21 falls within a small to medium range. This suggests that TOEFL scores could be associated with GPA. Effect size values, however, do not imply causation; that is, regardless of how strong the correlation between the variables of interest (i.e., TOEFL and GPA), having a higher TOEFL score does not lead to a higher GPA. In terms of clinical

significance, Durlak (2009) stated that an effect size value of .2 would be useful to admission policymakers when considering academic performance measures.

The average effect size value produced in this study for the association between students' TOEFL scores and GPA is comparable to values produced by other admission tests. In contrast to cognitive tests that typically have moderate correlations (.3 - .4) with measures of academic performance (Cohen, 1988), language tests usually have lower values, commonly below .3 (Davies, 1988). For example, the typical effect size value of the common cognitive tests on graduate GPA is .3 for the GRE; .4 for the GMAT; and .5 for the MCAT (Kuncel & Hezlett, 2007).

The effect size criteria discussed above, although useful, should be interpreted with caution as interpretation depends on the nature of study and research context (Ferguson, 2009). Therefore, researchers should interpret effect size values in relation to existing research findings rather than using specific criteria (Thompson, 2002). In comparison with results of previous meta-analyses, this study yielded an improvement on the predictive validity of TOEFL scores. Wongtrirat (2010) reported an average effect size of .187 between students' TOEFL scores and GPA. Yan (1994) reported a value of .3 between students' TOEFL scores and first-year GPA. As discussed previously, these reviews included studies examining only the older version of the TOEFL test. The reader should note that although the effect size value reported in this study is comparable to those reported in previous studies, this value is based on different analytic assumptions. Moreover, the current study adopted a broader scope of search and included a variety of contexts and test versions.

In judging the magnitude of the effect size for the current study, the reader should recall that an effect size value in a meta-analysis is the product of multiple studies with different sample sizes and research contexts. As a result, the effect size value in this study might not reflect the actual estimate. As previously discussed, other factors influence the predictive validity coefficients (e.g., criterion reliability and range restriction).

Do Effect Sizes vary across Studies?

Having combined the effect size values from included studies, the next goal was comparison of these values. The second question of the current study was answered by examining the consistency of effect size values across studies. The different graphical and statistical procedures revealed effect size values varied across studies. The values ranged from .0 to .7, suggesting noticeable large variation. Additionally, the results of heterogeneity tests indicated that the variation reflected a real difference among studies due to factors other than simple error.

Which of the following Factors Moderate the TOEFL-GPA Relationship: Publication Status, Degree Level, Test Version, and Research Setting?

As heterogeneity across effect size values was noticeable, a moderator analysis was conducted to detect potential variables contributing to variation. A subgroup analysis, using the analog to ANOVA procedure, was used to examine if the variation was due to real differences between studies. Of the four factors examined: (a) publication status, (b) graduate level, (c) test version, and (d) research setting; only the latter was found to be a moderator. In addition, studies conducted outside the U.S. were

found to have higher effect size values than those conducted in the U.S. However, the difference between international (n =11) and U.S studies (n = 36) should be interpreted with caution as the number of international studies was smaller. One possible cause for higher effect size values in international studies might be a publication bias toward language. Most of international studies in the current study were published studies with significant results. Non-English language studies are more likely to be published in English Journals when having significant results (Grégoire, Derderian, & Le Lorier, 1995).

Although the moderator analysis failed to reveal significant results for the remaining three moderators, a descriptive comparison of the effect size values for these moderators revealed useful information. This comparison revealed that the average effect size values were higher for (a) unpublished studies, (b) graduate level, and (c) the pBT version of the TOEFL. However, these differences were not statistically significant and should be interpreted with caution.

Implications

Considering the high stakes nature of the TOEFL test – the large number of international students taking the test and the universities use of TOEFL scores in decision making– the current study is of interest to both stakeholders. The findings of this study support the notion that international students' TOEFL scores are positively associated with GPA. However, these results do not mean TOEFL scores provide a definitive indicator for students' academic performance. A more appropriate

interpretation would be TOEFL scores are useful in making predictions about students' future academic performance.

Although results from the current study were not strong in terms of effect size values, the results were meaningful in terms of practical utility. Specifically, the results support the attitudes of some student and universities regarding the relative importance of TOEFL scores. Stakeholders (i.e., students, faculty, and admissions officers) in universities, therefore, should find these results useful in terms of guiding the academic performance of international students.

Implications Regarding Test Scores Use

The researcher, through findings of this study, drew the conclusion that international students' TOEFL scores explain a small amount of the variance in students' academic performance. TOEFL scores, therefore, might have better utility in conjunction with other factors, including: (a) personal, (b) cognitive, (c) socio-economic, (d) academic, and (e) professional. Although students' TOEFL scores can be useful in making inferences regarding students GPAs, the results from this study do not necessarily imply students with lower TOEFL scores are more likely to have lower GPAs. Indeed, Students with low or no TOEFL scores should not be denied admission based on their TOEFL scores. Therefore, TOEFL scores should not solely be relied upon in making predictions about students' academic performance.

Furthermore, being a valid predictor of academic performance does not denote that TOEFL scores are valid in all contexts and for all purposes. As previously discussed, validity concerns still exist regarding test scores-based interpretations and

uses (Messick, 1989) and caution should be used when making inference from these scores (Bachman, 2005). For the future, Test stakeholders should determine how TOEFL scores should be appropriately used in making admission decisions. These stakeholders should consider a variety of factors when using test scores including: (a) university admission policies, (b) academic disciplines, (c) admission requirements, and (d) applicant profiles.

One implication taken from this study relates to what extent TOEFL scores are linked to GPA. The present study provides insights about the contribution of TOEFL scores in students' academic performance. Currently, In light of the current study findings, TOEFL scores should be used as an incremental factor to identify students who are more likely to be successful rather than as a determining factor.

Another implication from the current study relates to how much weight TOEFL scores should be given. In other words, should the TOEFL scores be given more or less weight when using other factors (e.g., previous academic work, recommendation letters, and interviews) in admission decisions? Results from this study confirm the long held belief that the answer depends on individual cases for students, academic programs, and universities. Therefore, the researcher recommends individual universities conduct their own research to evaluate the use of TOEFL scores for academic purposes.

Accordingly, TOEFL scores should be used by university admissions officers in combination with information related to the overall quality of applicant profiles. In addition, university stakeholders should consider TOEFL section scores (i.e., reading, listening, speaking, and writing). Each TOEFL section contributes to the students' total

TOEFL scores (ETS, 2005). Finally, the proper use of TOEFL scores in making admission decisions should involve a shared judgment of: (a) policymakers, (b) applicants, (c) faculty, (d) admission officers, and (e) researchers (Wylie &Tannenbaum, 2006). Results of the current study provide information to each of these stakeholders in the process of identifying potential international students for admission.

Implication Regarding Moderators

The moderator analysis from the current study provides a foundation for further research to extend understanding of the predictive validity for TOEFL scores. This study was the first study in which a researcher identified and examined the relationship of moderator factors (i.e., test version and research setting) with the predictive validity of TOEFL scores.

One implication from the moderator analysis relates to TOEFL test versions and predictive validity. Specifically, as a result of the study findings, the researcher concludes that no test version is most effective. Therefore, universities and testing centers using older versions of the TOEFL test can use the findings of this meta-analysis to support their continued use of these versions (i.e., pBT). However, more research is needed on the newer TOEFL version to confirm this conclusion.

This study also has important implications regarding the research setting moderator. International studies (i.e., studies conducted outside the U.S.) had more predictive power than those conducted in the U.S. One implication of this finding is while the research setting should be considered in evaluating the test scores predictive validity, stakeholder and researchers should not consider the higher predictive validity

produced by international studies as a supporting evidence on the use of test scores in admission decisions.

The difference between international and U.S. studies might be due to other unexplored factors (e.g., sample size, range restriction, and application bias). However, the researcher, via the analyses, reached some conclusions about international studies. First, correlations from international studies are more variable than U.S. studies. Second, the results for the international setting might not be generalizable as the majority of international studies (a) use the pBT version (73%), (b) undergraduate level (73%), and (c) published studies (82%). Third, No generalization is to be made as the number of international studies is much smaller than those of U.S. More empirical research is needed to understand the factors leading to this difference in research setting.

Limitations

Despite contributions from the current study, a number of limitations exist. These limitations primarily pertain to the methodological aspects of the study and are common to all meta-analyses. First, all meta-analyses are subject to search bias due to deficiencies in the search process leading to incomplete study sets. Although the search process in the current study was comprehensive, it is not guaranteed that all relevant studies were located.

Second, all meta-analyses are at risk of selection bias resulting from restricted selection criteria. The researcher in current study attempted to identify the most relevant studies. However, using stringent search standards is more likely to exclude relevant studies and limit the power of meta-analysis to detect the effect of interest.

Third, meta-analyses are frequently praised as an objective research method. However, even with clearly specified coding instructions, the coding process is likely to be susceptible to coder's subjective judgment. Inconsistent coding can lead to missing relevant data such as moderators. Meta-analysts usually deal with coding bias by using multiple coders and then estimating coder agreement.

Fourth, meta-analyses are vulnerable to publication bias. While this bias might be due to search and selection biases discussed above, a major source of publication bias is the file-drawer problem (Rosenthal, 1979). This bias occurs when studies with statistically significant results are more likely to be published than studies with non-significant results. The presence of publication bias could harm the outcomes of the meta-analysis since it can overestimate the effect size. The current meta-analysis employed a number of procedures to detect and evaluate the effect of publication bias. Even though results were robust to publication bias, these results need to be interpreted with caution.

Fifth, some studies do not report the data required to conduct the meta-analysis. This lack of data leads to excluding potential studies. For example, some authors present their findings in terms of statistical significance without reporting results (i.e., correlation coefficients) permitting estimation of effect size. Moreover, other authors do not explicitly provide details about moderator variables. For example, in the current meta-analysis, some studies did not include which test version they used, preventing their use in the moderator analysis.

Finally, meta-analysis is sometimes criticized for mixing studies with different theoretical backgrounds and research designs. However, these differences enable researchers to investigate the interaction between study features (Glass, 2000). Following this view, the variation among studies was systematically modeled using moderator analysis.

Meta-analysis outcomes are valid as far as the quality of individual studies. One interesting finding of this meta-analysis that can have implication for the development of meta-analysis methods is that low quality studies (i.e., studies with small sample size) may overestimate the effect size estimate. However, more research is needed to confirm these results as some small studies might truly have sound methodological design that can detect larger effect size estimate. Some meta-analysts try to overcome the problem of study quality by coding for study quality. However, coding study quality can be problematic. One issue with this procedure concerns the criteria for judging the study quality. Judging for study quality might be challenging as there is no a set of standards for study quality that suit all contexts of meta-analyses.

In the current meta-analysis, it was not possible to code for the study quality since most of included studies were similar in their methodological features, they were observational, based on correlational data analysis and were retrospective studies with similar sampling designs. The only methodological aspect that could be employed for examining study quality is sample size. Yet, the impact of the sample size on the effect size estimate was examined under the publication bias analysis. Thus, the issue of study

quality does not seem to be a serious source of bias for the current meta-analysis outcomes.

Furthermore, moderator analysis can be utilized to give insights about the relationship between study quality and effect size estimate. For example, in this study, it was found that the average effect size is higher for international studies than for U.S. Studies. This difference in effect size between the two research contexts might be due to study quality. However, as discussed in the previous section, more research is needed to understand the factors leading to this difference in research setting.

Future Research

Recommendations for Future Research

The results from this meta-analysis provide a basis for future research. Guided by these results, researchers are encouraged to use additional studies, modify techniques, and incorporate contexts. Although the number of studies included in the study is acceptable, a large number of studies used the older version of the TOELF test (pBT). To have a more accurate picture of the TOEFL scores predictive validity, there is a need to conduct more studies on the iBT as the prevalent test version in use.

More empirical research is needed on the theoretical framework modeling the relationship between language proficiency and academic performance. This framework is important to help researchers identify valid indicators as well as relative predictors of academic performance. As explained in the literature review, academic performance is a multi-dimensional construct (O'Connor & Paunonen, 2007; Kuncel, 2003). Thus, rather than focusing on GPA as the sole indicator of academic performance, future research

should examine the predictive validity of TOEFL scores on a variety of factors, such as evaluations and degree completion.

Another area for future research would be a comparative analysis of the predictive power of other commonly used language tests, such as the IELTS. Using moderator analysis, the effect size estimates for the different tests could be compared.

Recommendations for Research Practices

Due to insufficient data in individual studies, it was not possible to examine several subgroups, such as (a) gender, (b) age, (c) country, and (d) language of the participants. In addition, a small number of studies included separate results for academic disciplines and TOEFL sections. Researchers, therefore, are encouraged to report separate results for these subgroups. Reporting these results for such variables could enable future researchers to evaluate the predictive power of TOEFL scores for these variables.

Researchers conducting predictive validity studies need to provide results in a manner that help stakeholders evaluate the practical utility of tests. These studies, therefore, should report effect size values and confidence intervals. In addition to providing statistical results, researchers need to provide qualitative interpretation of the effect size in the context of their respective areas to guide admission processing. In making admission decisions, stakeholders need to base judgments not only on the validity evidence of test scores but also on the consequences of test score use (e.g., educational, social, and economical).

Researchers should consider efficient methods in examining the TOEFL scores predictive validity. Since these scores reflect the intellectual development of students' performance, measures of this performance are more likely captured using longitudinal designs. These designs enable researchers to examine associations between TOEFL scores and academic performance using multiple variables, such as semester GPA and cumulative GPA.

Researchers should also employ mixed method designs in which the quantitative analysis of international students' TOEFL scores and GPA can be enhanced by gauging their perspectives on how language proficiency is related to academic performance. Using these designs, researchers should work to identify significant relationships between test scores and academic performance and also understand the processes that explain these relationships.

Although the current study employed a variety of statistical methods, including a random effects model, researchers attempting to replicate this research could use emerging multilevel model techniques. Both the fixed and random effects models, while the most common models used in meta-analysis, suffer from several weaknesses (Hox, 2010).

REFERENCES

Al-Ansari, S. & Al-Musawi, N. (2003). TOEFL and FCE tests as predictors of academic success for advanced students at the University of Bahrain. *Journal of Educational & Psychological Sciences, 4*(1), 7-24.

Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence. *Research in Higher Education*, *40*(4), 461-485.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association (APA). (2010). *Publication Manual of the American Psychological Association.* Washington, DC: American Psychological Association.

Annor, P. (2010). *Factors that affect the academic success of foreign students at Cardinal Stritch University* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://libezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/499871423?accountid=7082

Ayers, J. B., & Quattlebaum, R. F. (1992). TOEFL performance and success in a master's program in engineering. *Educational and Psychological Measurement, 52*(4), 973–975. doi:10.1177/0013164492052004021

Ayers, J. B., & Peters, R. M. (1977). Predictive validity of the Test of English as a Foreign Language for Asian graduate students in engineering, chemistry, or mathematics. *Educational and Psychological Measurement*, *37*(2), 461-463.

Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education*, *28*(1), 35-42.

Bachman, L. F. (1990). *Fundamental consideration in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

Becker, B.J. (2005). Failsafe *N* or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp.111-125). West Sussex, England: John Wiley & Sons.

Borenstein, M., Hedges, L.V., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-Analysis, Version 2*. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.

Bower, J. R., Conner, J. D., Gerritz, E. M., Jarmon, H., Sims, A. G., Tschan, R. E., Wilcox, L., & Vroman, C. (1971). *The AACRAO-AID participant selection and placement study: Report to the Office of International Training, Agency for International Development*. U.S. Department of State. Washington, D.C. Retrieved from http://pdf.usaid.gov/pdf_docs/PNAAS919.pdf

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, UK: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to testing language assessment*. New York, NY: McGraw-Hil.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Burmeister, J., McSpadden, E., Rakowski, J., Nalichowski, A., Yudelev, M., & Snyder, M. (2014). Correlation of admissions statistics to graduate student success in medical physics. *Journal of Applied Clinical Medical Physics*, *15*(1), 375-385.

Card, N. A. (2011). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, *28*(4), 523-539.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421-442.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach.* Thousand Oaks, CA: Sage.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language Proficiency and Academic Achievement*. (pp. 79–90). Clevedon, England: Multilingual Matters Ltd.

Daller, M. H., & Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, *4*(1), 173-193.

Davies, A. (1988). Operationalizing uncertainty in language testing: An argument in favour of content validity. *Language Testing*, *5*(1), 32-48.

de Souza, L. V. (2012). *Factors related to the acculturation stress of international students in a faith-based institution* (Doctoral dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/1071261790?accountid=7082

Des Brisay, M. (1994). Problems in developing an alternative to the TOEFL. *TESL Canada Journal*, *12*(1), 47-57.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917-928.

Duval, S., & Tweedie, R. (2000). Trim and Fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463.

Educational Testing Service (ETS). (1994). *Guidelines for the use of TOEFL scores, 1994–95 Edition*. Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-9495.pdf

Educational Testing Service (ETS). (1997). *TOEFL test and score manual*. Retrieved from http://t1t.net/book/save.php?action=saveattach&id=152

Educational Testing Service (ETS). (2001). *The Computer-based TOEFL score user guide*. Retrieved from http://www.emse.fr/~bsimon/documents%20p%E9dagogiques/p%E9dagogie/TOEIC%20TOEFL/989551.pdf

Educational Testing Service (ETS). (2005). *Setting the final cut scores*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/setting_final_scores.pdf

Educational Testing Service (ETS). (2007). *ETS test and score data summary for TOEFL CBT tests*. Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-0506-CBT.pdf

Educational Testing Service (ETS). (2011a). *TOEFL program history. TOEFL iBT Research Insight Series. 6*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf

Educational Testing Service (ETS). (2011b). *Test and score data summary for TOEFL® Internet-based and Paper-based test*. Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-2010.pdf

Educational Testing Service (ETS). (2011c). Reliability and comparability of TOEFL iBT scores. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Educational Testing Service (ETS). (2014). *Test and score data summary for TOEFL iBT tests*. Retrieved from http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ: British Medical Journal*, *315*(7109), 629-634.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing*, *2*(1), 1-17.

Fakeye, D., & Ogunsiji, Y. (2009). English language proficiency as a predictor of academic achievement among ELF students in Nigeria. *Journal of Science Research*, *37*(3), 490-495.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, *10*(3), 274-291.

Fenster, A., Markus, K. A., Wiedemann, C. F., Brackett, M. A., & Fernandez, J. (2001). Selecting tomorrow's forensic psychologists: A fresh look at some familiar predictors. *Educational and Psychological Measurement*, *61*(2), 336-348.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532-538.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think statistical. *Psychological Science*, *15*(2), 119-126.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. *Psychological Methods*, *6*(2), 161-180.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665-694.

Field, A. P. (2013). *Discovering Statistics using IBM SPSS Statistics* (2nd ed.). London, UK: SAGE Publications Ltd.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2-18.

Fu, Y. (2012). *The effectiveness of traditional admissions criteria in predicting college and graduate success for American and international students* (Doctoral dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/917742486?accountid=7082. (917742486)

Garg, A. X., Hackam, D., & Tonelli, M. (2008). Systematic review and meta-analysis: When one study is just not enough. *Clinical Journal of the American Society of Nephrology*, *3*(1), 253-260.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Glass, G.V. (2000). *Meta-analysis at 25*. Retrieved from http://www.gvglass.info/papers/meta25.html

Glass, G V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology.* Needham Heights, MA: Allyn & Bacon.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83-92.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, *74*(3), 249-266.

Gong, Y., & Fan, J. (2006). Longitudinal examination of the role of goal orientation in cross-cultural adjustment. Journal of Applied Psychology, 91(1), 176-184.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505-521.

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? *Journal of Clinical Epidemiology*, *48*(1), 159-163.

Hartnett, R. T., & Willingham, W. W. (1980). The criterion problem: What measure of success in graduate education? *Applied Psychological Measurement*, *4*(3), 281-291.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486-504.

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557-560.

Hill, K., Storch, N., & Lynch, B. (1999). *A comparison of IELTS and TOEFL as predictors of academic success*. IELTS Research Report. Retrieved from http://www.ielts.org/PDF/Hill,%20Storch,%20Lynch2.pdf

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hu, S. P. (1991*). English proficiency and academic performance of international graduate students* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/303911928?accountid=7082

Hughey, A. W., & Hinson, D. (1993). Assessing the efficacy of the Test of English as a Foreign Language. *Psychological Reports*, *73*(1), 187-193.

Hunter, J. E., & Schmidt, F. L. (2004*). Methods of meta-Analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Institute of International Education. (2013). *Open Doors Report on International Educational Exchange*. New York, NY: Institute of International of Education. Retrieved from http://www.iie.org/opendoors

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL® 2000 framework: A working paper* (TOEFL® Monograph No. MS-16). Retrieved from https://www.ets.org/Media/Research/pdf/RM-00-03.pdf

Johnson, P. (1988). English language proficiency and academic performance of undergraduate international students. *TESOL Quarterly*, *22*(1), 164-168.

Kamara, G. M. (1994). *Prediction of international teaching assistants' English proficiency and teaching effectiveness from panel examination and undergraduate students' evaluation* (Doctoral Dissertation). *ProQuest*

Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/304119575?accountid=7082

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64).Westport, CT: Greenwood Publishing.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2),137-152.

Kock, A., & Gemünden, H. G. (2009). A guideline to Meta-analysis. Retrieved from http://www.tim.tuberlin.de/fileadmin/fg101/TIM_Working_Paper_Series/Volume_2/TIM_WPS_Kock_2009.pdf

Kuncel, N. R. (2003). *The prediction and structure of academic performance*. Unpublished doctoral dissertation. Minneapolis, MN: University of Minnesota.

Kuncel, N.R., & Hezlett, S.A. (2007). Standardized tests predict graduate student's success. *Science*, *315*(5815), 1080-1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162-181.

Kwai, C. (. K.). K. (2010). *Model of international student persistence: Factors influencing retention of international undergraduate students at two public*

statewide four-year university systems (Doctoral dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305213878?accountid=7082. (305213878)

Lado, R., (1961). *Language testing: The construction and use of foreign language tests*. New York, NY: McGraw Hill.

Lei, P.-W., Bassiri, D., & Schulz, E. M. (2001). *Alternatives to the grade point average as a measure of academic achievement in college* (ACT Research Report Series 2001-4). Retrieved from http://files.eric.ed.gov/fulltext/ED462407.pdf

Liao, Y. F. (2004). Issues of validity and reliability in second language performance assessment. *Working Papers in TESOL & Applied Linguistics*, *4*(2), 1-4. Retrieved from http://www.tc.columbia.edu/academic/tesol/wjfiles/pdf/yenfenforum.pdf

Light, R. L., Xu, M., & Mossop, J. (1987). English proficiency and academic performance of international students. *TESOL Quarterly*, *21*(2), 251-261.

Lipsey, M.W. & Wilson, D.B. (2001). *Practical Meta-analysis.* Thousand Oaks, CA: Sage.

Lundahl, B., Yaffe, J., & Hobson, J. (2008). Today's studies, tomorrow's Meta-Analyses: Implications for evidence informed decision-making in social work. *Journal of Social Service Research*, *35*(1), 1-9.

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 (Master's thesis). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/895980467?accountid=7082. (895980467)

McMillan, J. H., & Foley, J. (2011). Reporting and discussing effect size: Still the road less traveled. *Practical Assessment, Research & Evaluation*, *16*(14), 1-12.

McNamara, T. F. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-104). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Messner, P.E., & Liu, N. (1995). The Test of English as a Foreign Language: An examination of "cut-off scores" in US universities. *The International Journal of Educational Management, 9*(2), 39-42.

Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. PLoS Med *6*(7), 1-6. doi:10.1371/journal.pmed.1000097

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355-371.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Neal, M. (1998). *The predictive validity of the GRE and TOEFL exams with GGPA as the criterion of graduate success for international graduate students in science and engineering*. Retrieved from http://files.eric.ed.gov/fulltext/ED424294.pdf

Nelson, C. V., Nelson, J. S., & Malone, B. G. (2004). Predicting success of international graduate students in an American university. *College and University, 80*(1), 19-27.

Ng, J. N. K. (2007). *Test of English as a Foreign Language (TOEFL): Good indicator for student success at community colleges?* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/304821394?accountid=7082. (304821394)

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (ACT Research Report Series 2002–4). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR2002-4.pdf

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971-990.

Palmer, L.A., & Woodford, P.E. (1978). English tests: Their credibility in foreign student admissions. *College and University, 53*, 500-510.

Pennock-Román, M. (2002). Relative effects of English proficiency on general admissions tests versus subject tests. *Research in Higher Education*, *43*(5), 601-623.

Perry, W. S. (1988). *The relationship of the Test of English as a Foreign Language (TOEFL) and other critical variables to the academic performance of international graduate students* (Doctoral dissertation). ProQuest Dissertations and These. Retrieved from http://libezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/303679735?accountid=7082. (303679735)

Person, N. E. (2002). *Assessment of TOEFL scores and ESL classes as criteria for admission to career & technical education and other selected Marshall University graduate programs* (Master's thesis). ProQuest Dissertations and These. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/62186537?accountid=7082. (62186537; ED473756)

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175-181.

Pitigoi-Aron, G., King, P. A., & Chambers, D. W. (2011). Predictors of academic performance for applicants to an international dental studies program in the United States. *Journal of Dental Education, 75*(12), 1577-1582.

Ren, J., & Hagedorn, L.S. (2012). International graduate students' academic performance: What are the influencing factors? *Journal of International Students, 2*(2), 135-143.

Richards, J. C. (2011). *Competence and performance in language teaching*. New York, NY: Cambridge University Press.

Rosenthal, R. (1979). The ‘‘file drawer problem’’ and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Sagan, C. (1997). *Demon-haunted world: Science as a candle in the dark*. London, UK: Headline book publishing

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section* (TOEFL iBT Research Report No. TOEFLiBT-08). Retrieved from http://www.ets.org/Media/Research/pdf/RR-09-02.pdf

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Report No. TOEFLiBT-04). Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-09.pdf

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*(4), 473-485.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274.

Schmidt, J. R. (1991). *The prediction of academic performance of Malaysian undergraduate students in an American-affiliated university program conducted in Malaysia* (Doctoral dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/303949067?accountid=7082

Schmit, S. E. (2001). *Locus of control orientation of international students within the Arizona community college system* (Doctoral dissertation). Dissertation Abstracts International Section A: Humanities and Social Sciences. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/619700841?accountid=7082. (619700841; 2001-95007-082)

Seaver, A. R. (2012). *Success of international students in higher education* (Master's thesis. Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=dayton1343416310&disposition=inline

Shohamy, E. (1996). Competence and performance in language testing. In G. Brown, K. Malmkaer, & J. Williams (Eds.), *Performance and competence in second*

language acquisition (pp. 138–151). Cambridge, UK: Cambridge University Press.

Simner, M. L. (1998). Use of the TOEFL as a standard for university admission: A position statement by the Canadian Psychological Association. *European Journal of Psychological Assessment*, *14*(3), 261-265.

Simner, M. L. (1999). Reply to the universities' reaction to the Canadian Psychological Association's position statement on the Test of English as a Foreign Language. *European Journal of Psychological Assessment*, *15*(3), 284-294.

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046-1055.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401-426.

Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing*, *21*(2), 146-173.

Stoynoff, S. (1997). Factors associated with international students' academic achievement. *Journal of Instructional Psychology, 24*(1), 56-69.

Tan, X., & Michel, R. (2011). Why do standardized testing programs report scaled scores? *ETS R & D Connections*, *16*, 1-6. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections16.pdf

The Organization for Economic Co-operation and Development. (2012). *Education at a Glance 2012: Highlights*. OECD Publishing. Retrieved from http://dx.doi.org/10.1787/ eag_highlights-2012-en

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology*, *133*(2), 133-140.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25-32.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J.W. Osborne (E.d.), *Best practices in quantitative methods* (246–262). Thousand Oaks, CA: Sage.

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analyses be trusted? *The Lancet*, *338*(8775), 1127-1130.

Torres, H., & Zeidler, D. (2002). The effects of English language proficiency and scientific reasoning skills on the acquisition of science content knowledge by Hispanic English language learners and native English language speaking students. *Electronic Journal of Science Education*, *6*(3). Retrieved from http://wolfweb.unr.edu/homepage/crowther/ejse/torreszeidler.pdf

Toulmin, S. E. (1958,). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Vinke, A. A., & Jochems, W. M. G. (1993). English proficiency and academic success in international postgraduate education. *Higher Education*, *26*(3), 275-285.

Vinz, S. (2012). Playing the TOEFL Game at SCSU. *Linguistic Portfolios*. Retrieved from http://repository.stcloudstate.edu/stcloud_ling/vol1/iss1/17

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. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education*, *98*(4), 389-398.

Wang, J. (2003). *A study of the adjustment of international graduate students at American universities, including both resilience characteristics and traditional background factors* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305326526?accountid=7082. (305326526)

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 259–318). New York, NY: Routledge.

Wilkinson, L., & Taskforce on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist*, *54*(8), 594–604.

Witt, A. S. (2010). *Establishing the validity of the task-based English speaking test (TBEST) for international teaching assistants* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305184053?accountid=7082. (305184053)

Wongtrirat, R. (2010). *English language proficiency and academic achievement of international students: A meta-analysis* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/744086570?accountid=7082. (744086570)

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.

Wylie, E. C. & Tannenbaum, R. J. (2006). *TOEFL® Academic speaking test: Setting a cut score for international teaching assistants* (ETS Research Memorandum). Retrieved from www.ets.org/Media/Tests/TOEFL/pdf/ngt_itastandards.pdf

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs* (TOEFL iBT Research Report No. TOEFLiBT-03). Retrieved from http://www.ets.org/Media/Research/pdf/RR-08-02.pdf.

Xu, M. (1991). The impact of English-language proficiency on international graduate students' perceived academic difficulty. *Research in Higher Education*, *32*(5), 557-570.

Yan, Z. (1994). *A meta-analysis on studies of TOEFL scores' predictive validity on first year's GPA* (1964 -1994). Unpublished manuscript, Department of Language Education, University of British Columbia, Vancouver, Canada.

Yan, Z. (1995). *Predictive validity of TOEFL scores on first term's GPA as the criterion for international exchange students* (Master's thesis). Retrieved from https://circle.ubc.ca/bitstream/handle/2429/3901/ubc_1995-0433.pdf

Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report). Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2001-6-differential-validity-prediction-college-admission-testing-review.pdf

Yule, G., & Hoffman, P. (1990). Predicting success for international teaching assistants in a US university. *TESOL Quarterly*, *24*(2), 227-243.

Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success*. Annual Meeting of the American Educational Research Association, Vancouver, Canada. Retrieved from http://cae-dev.ayera.net/images/uploads/pdf/Comparing_Alternatives_in_the_Prediction_of_College_Success.pdf

Zhang, Y. (2012). *An examination of acculturative stress, perceived social support and depression among Chinese international students* (Master's thesis). Retrieved from http://surface.syr.edu/cgi/viewcontent.cgi?article=1002&context=cfs_thesis

APPENDIX A

INCLUDED STUDIES

Al-Ansari, S. & Al- Musawi, N. (2003). TOEFL and FCE tests as predictors of academic success for advanced students at the University of Bahrain. *Journal of Educational & Psychological Sciences, 4*(1), 7-24.

Arcuino, C. L. T. (2014). *The relationship between the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) scores and academic success of international master's students* (Doctoral Dissertation). Dissertation Abstracts International Section A: Humanities and Social Sciences. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/1519509016?accountid=7082. (1519509016; 2014-99070-161)

Burmeister, J., McSpadden, E., Rakowski, J., Nalichowski, A., Yudelev, M., & Snyder, M. (2014). Correlation of admissions statistics to graduate student success in medical physics. *Journal of Applied Clinical Medical Physics / American College of Medical Physics, 15*(1), 4451. doi:http://dx.doi.org/10.1120/jacmp.v15i1.4451

Carty, R. M., Moss, M. M., Al-Zayyer, W., Kowitlawakul, Y., & Arietti, L. (2007). Predictors of success for Saudi Arabian students enrolled in an accelerated baccalaureate degree program in nursing in the United States. *Journal of Professional Nursing: Official Journal of the American Association of Colleges*

of Nursing, *23*(5), 301-308. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/68331256?accountid=7082

Chang, T., & Agronow, S. (2009). *Predicting success of international undergraduate students from non-English speaking countries at American universities*. Paper presented at the AIR Forum. Atlanta, GA.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421-442.

Dunn, J. W. (2006). *Academic adjustment of Chinese graduate students in United States institutions of higher education* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305316628?accountid=7082. (305316628)

Elliott, M. J. (2011). *Academic predictors of national council licensure examination for registered nurses pass rates* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/884081723?accountid=7082. (884081723)

Fass-Holmes, B., & Vaughn, A. A. (2014). Are international undergraduates struggling academically? *Journal of International Students*, *4*(1), 60-73.

Fournier, S. M., & Ineson, E. M. (2013). English proficiency of non-native speakers as a predictor of first-year performance in undergraduate hospitality management courses. *Journal of Hospitality & Tourism Education*, *25*(2), 49-56.

Fu, Y. (2012). *The effectiveness of traditional admissions criteria in predicting college and graduate success for American and international students* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/917742486?accountid=7082. (917742486)

Gong, Y., & Fan, J. (2006). Longitudinal examination of the role of goal orientation in cross-cultural adjustment. Journal of Applied Psychology, 91(1), 176-184.

Hoefer, P., & Gould, J. (2000). Assessment of admission criteria for predicting students' academic performance in graduate business programs. *Journal of Education for Business*, *75*(4), 225-229.

Itaya, L. E., Chambers, D. W., & King, P. A. (2008). Analyzing the influence of admissions criteria and cultural norms on success in an international dental studies program. *Journal of Dental Education*, *72*(3), 317-328.

Koys, D. (2009). GMAT versus alternatives: Predictive validity evidence from Central Europe and the Middle East. *Journal of Education for Business*, *85*(3), 180-185.

Kwai, C. (.K.). K. (2010). *Model of international student persistence: Factors influencing retention of international undergraduate students at two public statewide four-year university systems* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-

ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305213878?accountid=7082. (305213878)

Lee, Y. (2005). *Construct validation of an integrated, process-oriented, and computerized English for academic purposes (EAP) placement test: A mixed method approach* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://libezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305003620?accountid=7082. (305003620)

Lo, J. (2002). *The relationship between TOEFL scores and first-year GPA: A study of freshmen international students attending Texas A&M university-Kingsville from 1996-2001* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/305440501?accountid=7082. (305440501)

Maleki, A., & Zangani, E. (2007). A survey on the relationship between English language proficiency and the academic achievement of Iranian EFL students. *Asian EFL Journal*, *9*(1), 86-96.

Melnick, G. A., Kaur, G., & Yu, J. (2011). Social integration and academic outcomes: The case of an international public policy and management program. *Journal of Public Affairs Education*, *17*(4), 569–584.

Nelson, C. V., Nelson, J. S., & Malone, B. G. (2004). Predicting success of international graduate students in an American university. *College and University, 80*(1), 19-27.

Ng, J. N. K. (2007). *Test of English as a Foreign Language (TOEFL): Good indicator for student success at community colleges?* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/304821394 4?accountid=7082. (304821394)

Person, N. E. (2002). *Assessment of TOEFL scores and ESL classes as criteria for admission to career & technical education and other selected Marshall University graduate programs* (Master's thesis). Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/62186537 ?accountid=7082. (62186537; ED473756)

Pitigoi-Aron, G., King, P. A., & Chambers, D. W. (2011). Predictors of academic performance for applicants to an international dental studies program in the United States. *Journal of Dental Education*, *75*(12), 1577-1582.

Poyrazli, S., Arbona, C., Bullington, R., & Pisecco, S. (2001). Adjustment issues of Turkish college students studying in the United States. *College Student Journal*. *35*(1), 52-62

Sahragard, R., Baharloo, A., & Soozandehfar, S. M. A. (2011). A closer look at the relationship between academic achievement and language proficiency among

Iranian EFL students. *Theory and Practice in Language Studies*, *1*(12), 1740-1748.

Sailor, P. (2011). *The relationship of TOEFL scores to first-term GPA among new UCB international freshmen and transfers, summer 2011 through spring 2013*. Retrieved from http://www.colorado.edu/pba/records/TOEFLStudyResults.docx.

Salinas, A. (2008). *A study of TOEFL scores and first year grade point average of Latin graduate international students at Texas A&M university-Kingsville from 2003--2005 as a predictor of academic success in English* (Doctoral Dissertation). Dissertation Abstracts International Section A: Humanities and Social Sciences. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/621718350?accountid=7082. (621718350; 2008-99011-390)

Seaver, A. R. (2012). *Success of international students in higher education* (Master's thesis). Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=dayton1343416310&disposition=inline

Simner, M. L., & Mitchell, J. B. (2007). Validation of the TOEFL as a Canadian university admissions requirement. *Canadian Journal of School Psychology*, *22*(2), 182-190.

Stacey, D. G., & Whittaker, J. M. (2005). Predicting academic performance and clinical competency for international dental students: Seeking the most efficient and effective measures. *Journal of Dental Education, 69*(2), 270-280.

Takagi, K. K. (2011). *Predicting academic success in a Japanese international university* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/873966873?accountid=7082. (873966873)

Theuri, P. M., & O'Neill, K. (2007, June 24-26). *A correlation analysis of English language proficiency and performance in content-area cognitive skills*. Paper presented at the Oxford Business & Economics Conference. Oxford University, UK. Retrieved from http://www.gcbe.us/2007_OBEC/data/Kate%20Oneill,%20Peter%20M.%20Theuri.doc

Viravaidya, K, Panyakul, W R, & Thonglek, S. (2007, March 19-23). *Relative significance of admission criteria for the chemical engineering practice school (ChEPS)*. Paper presented at the 10th UICEE annual conference on engineering education. Bangkok, Thailand. Retrieved from http://www.chemeng.kmutt.ac.th/cheps/10thUICEE_ChEPS_Admission_Criteria.pdf

Vu, L. T. (2011). *An analysis of international graduate students' TOEFL scores, GPA, and perceptions of English language difficulties* (Doctoral Dissertation). ProQuest Dissertations and Theses. Retrieved from http://lib-ezproxy.tamu.edu:2048/login?url=http://search.proquest.com/docview/902586515?accountid=7082. (902586515)

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. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education*, *98*(4), 389-398.

Wang, W. (2013). *Testing the validity of GRE scores on predicting graduate performance for engineering students* (Master's thesis). Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1193&context=cehsdiss

Ward, T., & Jacobs, J. (2014). *International freshman performance preparation and success at the University of Oregon*. Retrieved from http://oem.uoregon.edu/sites/oem1.uoregon.edu/files/UO_Intl_FR_Feb2014_v1.1.pdf

Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney Papers in TESOL*, *1*(1), 51-70.

Zhang, Y. (2012). *An examination of acculturative stress, perceived social support and depression among Chinese international students* (Master's thesis). Retrieved from http://surface.syr.edu/cgi/viewcontent.cgi?article=1002&context=cfs_thesis

APPENDIX B

CODING BOOK

Study Characteristics	
• **Study ID**	The author last name Attach the full text (in hyperlink)
• **Biographic reference**	The complete citation in APA
• **Publication Year**	Publication year **from 2000 -**
• **Publication type** Published Unpublished	• Journal article • Book/book chapter • Thesis or doctoral dissertation • Technical report • Conference paper • unpublished manuscript • Other
Sample Characteristics	
• **Sample Size** Total number of participants (*N*)	Include only the sample that used the TOEFL variable. There are some studies which include more than one variable other than the TOEFL and each one has different sample size.
Outcome Measures	
• **Type of measure**	How measures were *operationalized?*
GPA	Semester GPA First-year GPA Cumulative GPA

Moderators	
• **Test version**	• pBT • cBT • iBT
• **Degree Level**	• Undergraduate (UG) • Graduate (G) (Master, Ph.D.)
• **Setting of study**	• U.S. • International (the country)
Other Variables	
• **Academic Disciplines**	• Humanities • Sciences • Business • Social sciences
• **TOEFL Sections**	• Listening • Reading • Writing • Speaking
• **Demographic variables**	• Gender (Male (M)/ Female (F) • Country/ language • Age
Results	
• **Type of Statistical Analysis**	• Inferential statics (correlation, regression, *t*-test, ANOVA), • Descriptive statics (Means, SDs): • A measure of effect size (e.g., Eta squared), • p-value • Other type of statistical test that can be converted to effect size
• **Statistical Results**	• Correlation coefficient (r), • Regression coefficient (β), • *t*-Statistic, *F*-Statistic, *p*-values

APPENDIX C

FLOW DIAGRAM

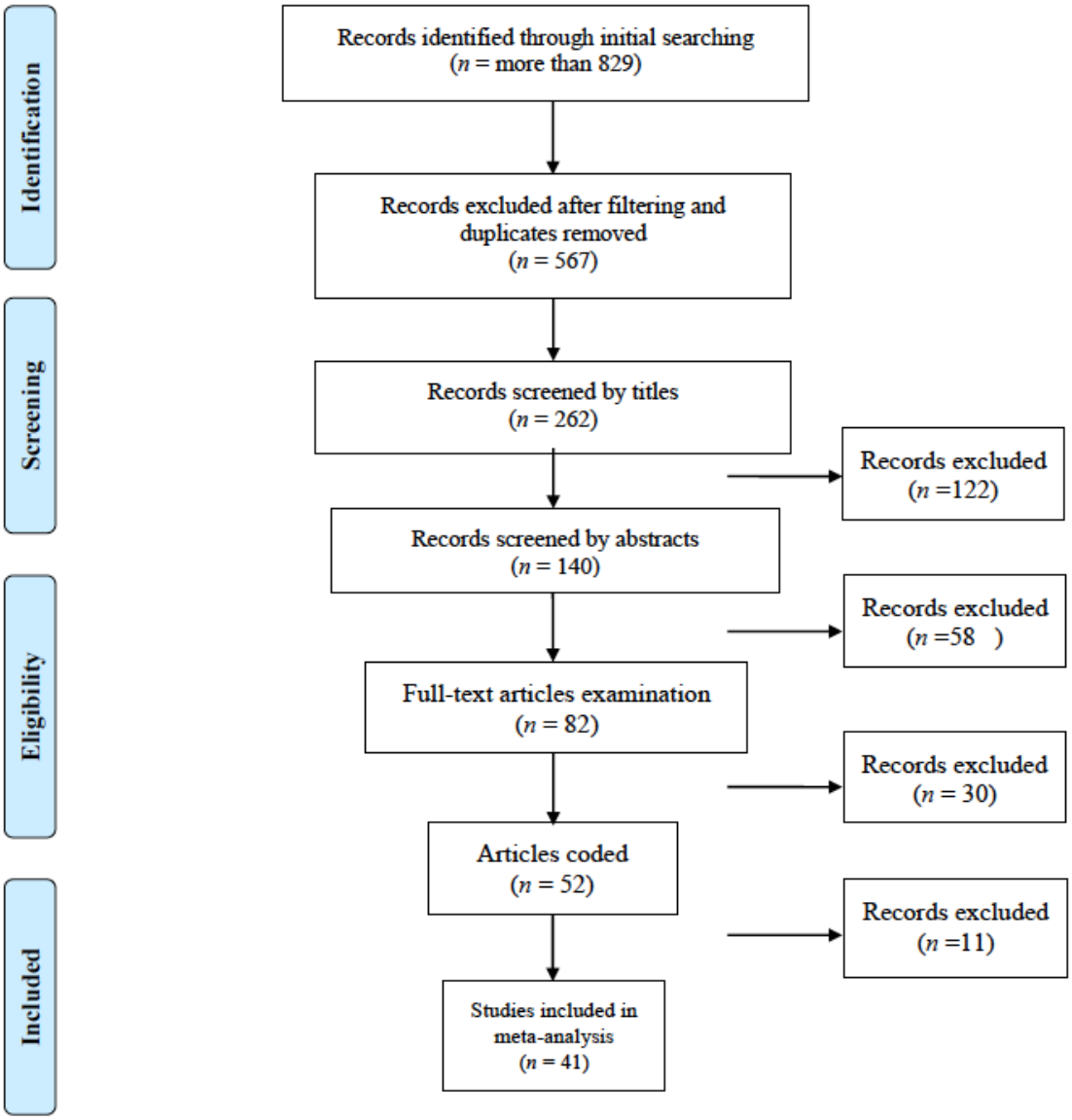

Chart adapted from Moher, Liberati, Tetzlaff, & Altman, The PRISMA Group (2009).