THE USE OF LANGUAGE PROFICIENCY TEST SCORES IN GRADUATE ADMISSIONS

by

Sharareh Taghizadeh Vahed

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of English

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair

Department of English

Dr. Tony Silva

Department of English

Dr. Irwin Weiser

Department of English

Dr. Xun Yan

Department of Linguistics

University of Illinois at Urbana-Champaign

Dr. Lixia Cheng

Purdue Language and Cultural Exchange (PLaCE)

Approved by:

Dr. Dorsey Armstrong

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Nadereh Sabokpey and Rasoul Taghizadeh Vahed who have been a constant source of support and encouragement during the challenges of graduate school and life. My sister Bahareh has never left my side and is very special. I dedicate this work and give special thanks to my husband and best friend Mehdi Marashi for being there for me throughout the entire doctorate program. I also dedicate this dissertation to my many friends who have supported me throughout the process. I will always appreciate all they have done.

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my advisor, Prof. April Ginther, for her invaluable advice, continuous support, and patience during my PhD study. Her immense knowledge and plentiful experience have encouraged me in all my academic research and daily life. I would also like to thank Dr. Tom Atkinson, Associate Dean of the Graduate School, for his help and advice, and Mr. Jeff Bridgham, Senior Data Analyst at Purdue's Graduate School, for providing me with the data I used for my dissertation analyses. I would like to express my sincere gratitude to my Dissertation Committee Members for reading this dissertation and helping me improve this work. My gratitude extends to the Faculty of English Department and the Oral English Proficiency Program for the funding opportunity to undertake my studies. I would like to thank all members of the SLS community here at Purdue. It is their kind help and support that have made my study and life at Purdue a wonderful time. Finally, I would like to express my gratitude to my parents and my husband. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

TABLE OF CONTENTS

LIST OF TABLES 7

LIST OF FIGURES 8

ABSTRACT 10

CHAPTER 1. INTRODUCTION 11

1.1 Problem Statement 11

1.2 Purpose of the Study 12

1.3 Organization of the Study 14

CHAPTER 2. LITERATURE REVIEW 16

2.1 Introduction 16

2.2 Review of the Related Literature 17

2.2.1 Graduate Admissions 17

2.2.2 Language Assessment Literacy (LAL) of Admissions Decision Makers 42

2.3 Conclusion 55

CHAPTER 3. RESEARCH METHODS 56

3.1 Purpose of Research 56

3.1.1 The Need for Language Proficiency Literacy 57

3.1.2 Interview with the Associate Dean of Graduate School 59

3.2 Data Analysis 61

3.2.1 Tabulation and Graphing of Descriptive Data 62

3.2.2 Distribution of TOEFL Scores by Admission Status 63

3.2.3 The Cluster Analysis Procedure 64

3.3 Conclusion 65

CHAPTER 4. RESULTS AND DISCUSSION 66

4.1 International Applicant Pool 66

4.2 Language Proficiency Profiles 73

4.2.1 Implications for Language Assessment Literacy (LAL) 86

4.3 TOEFL Score Distributions 91

4.3.1 Implications for Language Assessment Literacy 93

CHAPTER 5. CONCLUSIONS, LIMITATIONS, AND RECOMMENDATIONS FOR FURTHER RESEARCH 96

5.1 Summary of the Study Findings and Implications 96

5.2 Limitations of the Study 97

5.3 Recommendations for Further Research 98

APPENDIX A. OEPT HOLISTIC SCALE 100

APPENDIX B. TOEFL iBT SCORE USE AND INTERPRETATION MEMO, 2020 102

REFERENCES 104

LIST OF TABLES

Table 2.1. TOEFL Reliability Estimates 23

Table 2.2. Common Reference levels in CEFR 33

Table 2.3. TOEFL iBT Test Score Requirements in Relation to the Mapping of TOEFL iBT Test Scores at the B2 Level of the CEFR 35

Table 4.1. Full-time Graduate Students and the Percent of International Students by Field (2010) 68

Table 4.2.Top Ten Countries of Origin for International Students (2015/16) 74

Table 4.3. Top Ten Countries that Represent the Student Body at Purdue University in 2018 ... 75

Table 4.4. Cluster Centroids for Subscale Score Profiles 81

Table 4.5. Chi-square Test of Independence between Language Background and TOEFL Profile Membership 81

Table 4.6. Five Stages of Literacy in LAL 86

Table 4.7. Five Stages of Literacy in LPL 87

LIST OF FIGURES

Figure 2.1. Elements and context of an interpretive argument 28

Figure 2.2. The Percentage of Students in Five different TOEFL Speaking Score Categories in Each OEPT Score Level (N=1016) 39

Figure 2.3. TOEFL iBT/Common European Framework of Reference (CEFR) descriptors for participation in university activities 41

Figure 2.4. Levels of Language Proficiency Test Score Use in the Context of Graduate Admissions 46

Figure 2.5. The Balance between Language Proficiency Test Scores and Other Factors Important in Graduate Admissions Decisions 48

Figure 3.1. A Hypothetical TOEFL Subskill Score Distribution of Admitted and Rejected Applicants 64

Figure 4.1. Number of All, Admitted, and Matriculated International and Domestic Applicants in College of Engineering – AY 2018/19 69

Figure 4.2. Number of International and Domestic Applicants by Admission Status in College of Liberal Arts – AY 2017/18, 18/19, FALL 2019 70

Figure 4.3. Number of All International and Domestic Applicants in College of Engineering Academic Departments – AY 2018/19 71

Figure 4.4. Number of Admitted International and Domestic Applicants in College of Engineering Academic Departments – AY 2018/19 71

Figure 4.5. Number of All International and Domestic Applicants in College of Liberal Arts Academic Departments – AY 2017/18, 18/19, FALL 2019 72

Figure 4.6. Number of Admitted International and Domestic Applicants in College of Liberal Arts Academic Departments – AY 2017/18, 18/19, FALL 2019 73

Figure 4.7. Language Backgrounds of Admitted Applicants – College of Science 2018/19 75

Figure 4.8. Language Backgrounds of Admitted Applicants – College of Engineering 2018/19 76

Figure 4.9. Scree Plot for the Change in Agglomeration Coefficients in Purdue's Admitted Applicants' Language Proficiency Profiles – College of Engineering AY 2018/19 80

Figure 4.10. TOEFL Speaking Score Distribution across Language Background – College of Engineering AY 2018/19 82

Figure 4.11. TOEFL Reading Score Distribution across Language Background – College of Engineering AY 2018/19 83

Figure 4.12. The Percentage of Students in Five different TOEFL Speaking Score Categories in Each OEPT Score Level 90

Figure 4.13. TOEFL Speaking Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19 92

Figure 4.14. TOEFL Reading Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19 92

Figure 4.15. TOEFL Listening Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19 93

Figure 4.16. TOEFL Writing Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19 93

Figure 4.17. A Hypothetical TOEFL Subskill Score Distribution of Admitted and Rejected Applicants in College of Engineering 95

ABSTRACT

The purpose of this research is to reveal and compare the language proficiency profiles of Purdue's Chinese and Indian graduate applicants in various disciplines to take a step towards the development of Language Proficiency Literacy (LPL) of graduate admissions decision makers. The study argues that before being able to offer LPL development opportunities to admissions decision-makers, language testers need to gain admissions literacy in their specific academic context. One way this can be achieved is by analyzing graduate admissions data to see patterns of test score use in each discipline and to reveal language proficiency profiles of graduate applicants. Providing admissions decision makers with information about the linguistic characteristics of their applicants can be a very helpful step towards enhancing LPL in the context of graduate admissions.

One of the analyses conducted towards the goal LPL development in the context of graduate admissions was a Cluster Analysis procedure followed by a Chi-square analysis to compare the language proficiency profiles of graduate applicants from various L1 backgrounds based on scores on the Test of English as a Foreign Language (TOEFL). The study found three language proficiency profiles in graduate applicants' TOEFL data: 1) the 'unbalanced' profile, which consists of applicants who have higher scores in the subskills of reading and listening, and comparatively lower scores on speaking and writing, 2) the 'balanced medium' profile, which represents students who have moderate scores across all four subskills, and 3) the 'balanced high' profile, which consists of applicants who have high scores across all four subskills. The study found evidence for the interaction between graduate applicant test-takers' L1 background and belonging to a balanced or an unbalanced language proficiency profile, which highlights the importance of considering subskill scores in addition the total score when using language proficiency test scores to select graduate students from specific L1 backgrounds.

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

In the admission of graduate applicants to graduate programs, the use of standardized test scores is considered useful for cutting through the ambiguities involved in the process of student selection (Posselt, 2016). Standardized language test scores are one form of quantification used by admissions committees to select graduate students. Sufficient understanding of both the interpretations and limitations of standardized language test score use is necessary to make informed admissions decisions. Therefore, studying the use of language test scores by graduate admissions committees in the high-stakes enterprise of graduate admissions helps us have a better understanding of how language testers can contribute to increasing Language Proficiency Literacy (LPL) of the decision-making faculty.

Pursuing graduate-level education provides the opportunity for students to gain specialized knowledge beyond what is learned at the undergraduate level. Although graduate schools seek to admit candidates who are already specialists in their disciplines, it is usually a combination of multiple factors that leads to an admission offer. Among the most common required graduate application materials are standardized test scores, field-related work experience, field-related education, recommendation letters, writing samples, a minimum GPA, and a purpose statement. Posselt (2016) states that admissions decision-makers express concern about the use of "explicit cutoffs or tacit minima" when considering the weight of standardized test scores in student admissions. The rise in the volume of graduate applications received by higher education institutions leaves admissions committees wondering how excluding applicants based on their meeting a standardized test cut score affects the selection of students who nonetheless may be academically successful.

Due to the impracticality of sending out a large number of applications to the decision-making faculty, graduate schools set quantifiable standards at an initial stage during the process of admissions. "Putting numbers to judgments in order to simplify comparisons among applicants" is central to the graduate admissions review process (Posselt, 2016, p.30). Therefore, it is important to investigate how standardized test scores are implemented by the graduate admissions decision-makers after graduate applications leave the graduate school. Language

proficiency is one student selection criterion that has often baffled admissions decision-makers. To investigate the use of language proficiency test scores in the process of admissions, Ginther & Elder (2014) collected data using a survey and post-survey interviews and reported that while most of their respondents were aware of the level of English at which their students were admitted, many had misconceptions or uncertainties about what the university cut scores for English proficiency really represented. These uncertainties can influence the ultimate weight that standardized test scores carry in the process of admissions through either overreliance or under-reliance on cut scores set by the graduate school. Therefore, it is important to find out about the misconceptions and uncertainties and address them to ensure that score users are using language proficiency test scores in an informed way.

Language assessment literacy (LAL), which is defined as the knowledge that stakeholders involved in the assessment of languages are required to have, has been only scantly addressed in the context of graduate admissions research. O'Loughlin (2013) states that LAL "includes the acquisition of a range of skills related to test production, test score interpretation and use, and test evaluation in conjunction with the development of a critical understanding about the roles and functions of assessment within education and society" (p.363). While there is a considerably large body of literature about the importance of LAL training among language teachers and language practitioners, the important role of LAL among admissions decision-makers has been under-researched. This research, therefore, aims to investigate the use of language proficiency test scores by graduate admissions decision-makers as one of the main stakeholders in the process of assessing academic language proficiency.

1.2 Purpose of the Study

University admissions committees are among the key stakeholders in the use of language assessment results (Baker 2016), yet they are the most neglected group in terms of language assessment literacy development (O'Loughlin, 2013). High stakes decisions made by this group of test users necessitates attention to how they use language proficiency test scores for admissions purposes, and how LAL development, redefined in the context of graduate admissions, can be provided by language testers. Ginther & Elder (2014) state three reasons for the essential of admissions committee LAL research:

> (a) The increasing numbers of international students undertaking graduate level study in institutions of higher learning around the world, (b) the importance of English as a vehicle of communication in an increasingly global society, and (c) growing concerns … about the limited capacity of many international students to participate effectively in their study programs because of limited English proficiency (p. 2).

Ginther & Elder's (2014) research on admissions committees' LAL at Purdue University and the University of Melbourne revealed that there were "significant numbers [of admissions test score users] erroneously believing that the current minimum requirements indicated more than just an acceptable level of English" (p. 22) and that "assessment literacy among the [survey] respondents was generally limited" (p. 26). Their results also revealed that "many of the study's participants … expressed interest in further information" about language proficiency test scores and the minimum cutoffs (p. 26). The fact that this group of stakeholders acknowledge a gap in their knowledge about language proficiency test scores indicates the importance of our roles as language testing practitioners to provide the necessary information and try to fill the existing gap.

The purpose of the proposed study is to investigate the language assessment practices of the graduate admissions committees in various colleges, departments, and programs at Purdue University to reveal preferences and score use patterns when selecting students using language proficiency as a selection criterion. An important step towards the goal of providing admissions language test score users with the necessary information is to understand (a) how language test scores are currently being used by the decision makers, (b) what kind of interaction exists between English proficiency test score use and the users' academic discipline, and (c) how language test scores are being used across other factors such as applicants' language backgrounds, final admission status, matriculation status, and language proficiency profiles. Therefore, the proposed research will address the following research questions:

1. What are the characteristics of Purdue's graduate applicant pool in terms of language proficiency test score distribution across admission and matriculation status?
2. How do the distributions of total and subskill TOEFL scores compare across the two major language backgrounds (Indian and Chinese) of admitted, rejected, and matriculated applicants?

3. What are the language proficiency profiles of admitted graduate applicants and is there an association between proficiency profile membership and applicants' language backgrounds?

The findings of the present study will guide our attempts to help the decision-making faculty increase their LPL. The study will also benefit the Oral English Proficiency Program (OEPP) because Purdue's selectivity and graduate student selection dynamics affect the OEPP international teaching assistant (ITA) verification practices. According to Ginther (2003), in the 1980s, "the strength of the public perception of undergraduate difficulties with ITAs led to the establishment of mandates … requiring that the oral English proficiency of prospective ITAs be certified before they would be allowed to have direct contact with undergraduates" (p. 59). Purdue's OEPP was established in 1987 to fulfill the Purdue University requirement which states that any student whose first language is not English, and who is to be appointed as a teaching assistant, must demonstrate sufficient oral English proficiency before beginning their appointment and before they have direct contact with undergraduate students. The OEPP uses the Oral English Proficiency Test (OEPT), a locally designed computer-delivered oral English proficiency test, to certify students and exempt them from enrollment in OEPP's ENGL 620 course, Classroom Communication for International Graduate Students. The number of students requiring post-entry assessment of language proficiency and the number of those requiring post-entry language support through ENGL 620 are greatly affected by how the decision-making faculty use language proficiency test-scores in the admissions process. Therefore, taking steps in understanding graduate decision-makers' test score use patterns and addressing misconceptions about language proficiency are important in keeping post-entry language assessment and support at a manageable level.

1.3 Organization of the Study

This dissertation consists of five chapters. The present chapter, chapter one, introduces the background and the motivations to conduct the present research. Chapter 2 reviews the literature related to the use of language proficiency test scores in university admissions and its relevance to language assessment literacy. The chapter also provides a thorough review of the graduate admissions models and how policy can limit the complete operationalization of those models. In

addition, chapter 2 discusses the validity and reliability of the Test of English as a Foreign Language (TOEFL) as the most prevalently used test for graduate admissions. Last but not least, the chapter discusses the operationalization of language assessment literacy development in the context of higher education to present what is missing and what is necessary to address. Chapter 3 provides detailed information regarding the analysis of Purdue's graduate admissions data to reveal patterns of score use by the graduate admissions decision-makers. Chapter 4 reports the findings of the study based on the results obtained from the analysis of graduate admissions data and discusses the results based on the relevant literature. The last chapter, chapter 5, concludes the study and discusses limitations. This chapter also offers implications for the findings and insight into the directions of further research related to the use of test scores in graduate admissions.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

Admission of international graduate students to institutions of higher education in the US is a high-stakes enterprise for all involved: for applicants (because of their present and future employment opportunities), for admissions committees (because of their desire to enhance the quality of their programs and to complete funded research), and for the university enrollment managers because of their desire to maintain or increase enrollment and associated revenue. Standardized language test scores are used as a criterion in the admission of international students. Meeting the standard of appropriate score use and interpretation in the process of university admissions requires sufficient understanding of both the use and interpretation of standardized language test scores. Studying the use of language test scores by the graduate admissions committees and the extent to which policies limit their selection will help us have a better understanding of how language assessment community can contribute more effectively to the selection process in the realm of graduate admissions.

The purpose of this research is to investigate the patterns of language test score use by graduate admissions committees in various colleges and departments at Purdue University to reveal their preferences and practices when selecting students using language proficiency as a selection criterion. This chapter presents a review of the literature related to the use of language proficiency test scores in admissions and the language assessment literacy necessary to be able to use language test scores to make informed decisions during the admissions process. The chapter presents a review of the history and dynamics of graduate admissions in North America and discusses different models used in graduate admissions, the role of language proficiency scores in graduate admissions and international teaching assistant certification, the validity and reliability of the commonly-used language proficiency tests for admissions, the relationship between the Test of English as a Foreign Language (TOEFL) and the Common European Framework of Reference (CEFR), and the importance of language assessment literacy in the use of language proficiency test scores in graduate admissions. The chapter ends with a discussion about the term ‘language assessment literacy’ in the context of graduate admissions.

2.2 Review of the Related Literature

2.2.1 Graduate Admissions

History and Structure

The history of graduate education in the U.S. dates back to 1861 when the first doctoral degree was earned by a graduate student at Yale University after two years of post-graduate work away from the Yale campus (Hollis, 1945). It was not until the 1930s that Harvard University expressed concerns about the admission process and the need for the process to be standardized. In the Annual Report of the President published by Harvard University, President Lowell stated that the development of a set of standards that would encourage the best undergraduate students to aspire to earn a Doctor of Philosophy degree was important (Lowell, 1932). According to Hollis (1945), one of the reasons graduate education grew in the United States was because students were "flocking to European universities, especially to those in Germany", and as the first graduate school in the United States, Johns Hopkins graduate programs were established "to compete with these universities by reproducing most of their good characteristics and at the same time serving contemporary needs in the United States to a degree not possible for a foreign institution" (Hollis, 1945, p. 358). Becoming a world leader across every sphere of human activity, business, technology, science, politics, media, and of course education, the U.S. soon was home to many of the world's most prestigious graduate programs that drew international students from all over the world.

Graduate education experienced a significant growth in the 1950s, and with the growth came reviews of graduate admissions practices in various fields (Michel, Belur, Naemi, & Kell, 2019). Prior to World War II, only a small number of white males continued to higher education. The expansion of higher education in North America happened after World War II and the Korean War when the Servicemen's Readjustment Act of 1944, commonly known as the G.I. Bill, provided funding for returning soldiers' and their families' higher education (Gumport, Iannozzi, Shaman, & Zemsky, 1997). Other factors that contributed to the expansion of higher education in 1950's was the expansion of the middle class, the rapid development of suburban areas, and increasing family wealth. Families who desired more social and economic mobility began enrolling their children in higher education programs. Between 1970 and 1976, there was an expansion in the number of academic institutions that began offering master's and doctoral

degrees, while baccalaureate-granting institutions decreased in number (Gumport et al., 1997). With the expansion of graduate studies, extensive research began to examine the procedures for graduate student selection in higher education institutes (e.g. Burns, 1970; Carmichael, 1961; Harmon, 1966; Schwager, Hülsheger, Bridgeman, & Lang, 2015; Willingham, 1974), followed by more field-specific selection procedure review and research (e.g. Hall, O'Connell, & Cook, 2017; King, Bruce, & Gilligan, 1993; Mamary & Roe, 2004; Marks, 2011; Pitcher & Schrader,1972; Rock, 1974).

The process of graduate admissions in almost all graduate schools in the United States is decentralized as compared to undergraduate admissions, meaning that student selection decisions are made at the department- and program-level rather than by a centralized enrollment management office (Kent & McCarthy, 2016; Michel et al., 2019). The role of graduate schools in universities is partly administrative, i.e., checking the application materials to ensure that the applications are complete before sending them to departments. In Kent & McCarthy's (2016) study, which was conducted with 857 individuals in 250 institutions, more than 75% of graduate school staff indicated that academic departments were primarily responsible for graduate student selection, while less than 15% indicated that the graduate school was responsible for this task. These numbers indicate the decentralized nature of graduate admissions in many schools in North America. In an interview with Dr. Thomas Atkinson, the Associate Dean of the Graduate School at Purdue University, he explained that admissions at Purdue University is also decentralized, relying on departmental and program-level committees for decision making.

"Recruitment, admissions, and support in doctoral programs are often done by part-time committees of busy researchers and teachers and by individual faculty members, making it much more difficult to monitor or intervene in these processes" (Orfield, 2014, p. 453). Although the process of graduate admissions is usually decentralized, the graduate school staff, or in some cases, the dean of graduate school are responsible for making the final decision which usually aligns with the academic department's decision unless there is something wrong with a specific applicant's application materials (Michel et al., 2019). However, sometimes, the graduate school might have some goals (e.g. diversity) that may not necessarily match the goals of academic departments (e.g. research background, test scores, etc.) during the graduate admissions process. According to Orfield (2014), one problem with a decentralized admissions process is "the limited and confusing legal framework guiding the affirmative action process, some of which does not

fit well with the perspectives of faculty members in charge of the graduate admissions process" (p. 453). "Faculty motivations for considering diversity may or may not reflect institutional priorities and current policies" (Orfield, 2014, p. 453). Such disagreement between the goals of the graduate school and specific departmental needs and priorities creates challenges when it come to the implementation of guidelines provided by graduate school for graduate student selection.

Despite the decentralized nature of the graduate admissions, a specific set of application materials are usually required of applicants to be submitted to the graduate schools in various academic institutions. These materials can be classified into two major categories: 1. measures of cognitive skills, 2. measures of behavioral skills. During the graduate admissions process, cognitive skills, which are more tangible, are measured using two sets of application materials: undergraduate transcripts/GPA and scores on standardized tests (e.g. TOEFL, GRE, GMAT). Depending on the program, research background and academic publications might also play a role in indicating applicants' potential for success. Behavioral skills such as persistence and commitment are measured less objectively, using personal statement letters, recommendation letters, interviews, etc. Some departments require students also to submit diversity statements to examine eligibility for targeted support (diversity fellowships). Although behavioral skills can be as important as cognitive skills for graduate student success, standardized test scores are more frequently researched since they lend themselves to statistical analysis The predictive validity of the GRE and the TOEFL has been researched for undergraduate and graduate students in numerous studies using factors such as first-year GPA (e.g. Cho & Bridgeman, 2012; Ginther & Yan, 2018; Kuncel, Hezlett, & Ones, 2001; Kuncel, Wee, Serafin, & Hezlett, 2010), faculty ratings of performance (e.g. Reilly, 1976), and other factors such as number of publications, time to degree, and publications citations. Although ETS has consistently cautioned against the use of a single criterion as the sole predictor of students' future academic performance, standardized test scores and cut scores set by schools and departments are widely used by admissions decision makers (Michel et al., 2019).

Graduate Admission Models

'Holistic file review' is the most common method of graduate student admission, a process through which admissions decision makers can take both cognitive and behavioral skills of the

applicants into consideration. One of the major goals of implementing holistic review is its contribution to diversity and the admission of underrepresented minorities. Kent & McCarthy (2016) state in their comprehensive study on the holistic file review method that a truly holistic method "holds out great promise as a strategy for addressing issues of access and diversity" (p. 1). The holistic file review method can contribute to both underrepresented minority diversity and international student diversity. While international students' economic contributions to universities tend to be a primary focus, the cultural, political, and historical perspectives that help build vibrant, diverse campuses are also important for universities. Kent & McCarthy (2016) state that there is evidence to believe "holistic approaches result in similar or improved institutional performance on student success measures" (p. 1). Despite the merits of the holistic file review, there are challenges and complications this method can bring to the process of graduate admissions which might lead to reluctance in truly putting into practice a holistic review method.

Holistic file review can be more time consuming than methods that weigh standardized test scores. Some institutions still use cut scores to decrease the pool of applicants to be considered for admission in the holistic file review. This poses the problem of not being able to implement the holistic review method in its true sense, which is the review of each and every application as a whole regardless of strengths and weaknesses in each individual application material. That being said, some initial screening of English proficiency using minimum language proficiency test scores is important and necessary, especially given the limited resources for post-entry language support.

In the holistic review method, it can be difficult to assess behavioral skills as tangibly as cognitive skills are assessed. Current methods used to measure behavioral skills, (e.g., letters of recommendation, purpose statements, personal statements, etc.) are highly subjective. Letters of recommendation are almost always positive as recommenders are selected by the applicants (Michel et al. 2019). GlenMaye & Oakes (2002) asked 12 faculty members in the Social Work master's program to assign objective scores to a group of applicants' personal statements and found that the inter-rater reliability coefficient for these ratings was .50, which is below the .70 recommended threshold. The results of this research indicate the difficulty of reaching an agreement about what behavioral skills are more important for graduate student success than others.

As mentioned before, one of the major issues with holistic file review is the difficulty in its implementation with the large number of graduate applications universities need to consider for admission each year. In some cases, the graduate school uses standardized test scores and GPA in the application files to maintain a minimum standard which, in turn, decreases the number of applications sent to each department in an effort to make the holistic review process more manageable. However, this practice might not always align with the objectives of holistic file review and might affect the diversity goals of the university (Michel et al., 2019; Posselt, 2016). To address this challenge, some departments "triage" the applications, ranking them based on scores on standardized tests rather than eliminating some applicants altogether. Another way to ensure personal faculty biases are not affecting the decision-making process is to admit cohorts of students who find the mentors that best fit their academic and research interests several semesters after they start their graduate studies (Michel et al., 2019).

Language Proficiency Assessment in Higher Education

Many higher education institutions around the world require evidence of appropriate entry-level of English proficiency to ensure their prospective students have the ability to cope with the linguistic demands of graduate studies. The earliest test intended to specifically assess test-takers' English proficiency in the U.S. was the English Competence Examination. This test was developed by the College Entrance Examination Board in 1930 after a rapid growth in the number of international students applying to U.S. universities. The development of the test was also the result of a memorandum issued by the Commissioner General of Immigration that educational institutions need to ensure sufficient English proficiency before admitting international students (Spolsky, 1995).

In line with the demand for academic language assessment, other language testing agencies were founded too and began to design tests that would assess English in academic settings (Elder, 2017; Spolsky, 1995). By 1960, English language testing in the U.S. reached a high level of sophistication, and in 1961, ETS started to develop an English proficiency test to replace the five-hour *English Examination for Foreign Students* administered by the College Board to admit students to U.S. schools at the time (Spolsky, 1995) which resulted in the development of the first of three versions of the TOEFL. Two of the most frequently used tests for admissions purposes now are the Test of English as a Foreign Language (TOEFL iBT) and the International

English Language Testing Service (IELTS Academic). With TOEFL being the most prevalent in North America, the focus of this literature review will mainly be on this test.

The reliability of the TOEFL

Reliability refers to the degree to which scores obtained from a test represent test-takers' true scores based on classical test theory. Theoretically speaking, reliability is the relationship (correlation) between a person's score on equivalent forms (test-retest) of a test or across various items within a test (internal consistency). One of the major sources of inconsistency in a test is random measurement error. As more error is associated with the observed score, the lower the reliability will be. As measurement error decreases, reliability increases. Typically, threats to reliability include any variable other than the language ability being measured that affects test scores. If the error of measurement is high, it can limit the generalizability of test scores, meaning that it cannot be ensured that the results obtained will be similar if the test is taken several times by the same test-taker. There are methods of measuring reliability that can measure the random error of measurement across equivalent forms of a test or multiple administrations of a test and methods that measure error across items within a test. The *Test-retest* method measures the ability of a test to consistently measure a construct over a period of time. Fluctuations in test-takers' observed scores from one test administration to another determine the test-retest reliability of a test (Crocker & Algina, 1986).

Internal reliability, on the other hand, assesses the consistency of results across items within a test. Measuring the *Internal consistency* of a test performance is one of the most frequently used internal reliability measurements, which measures whether test items that propose to measure the same general construct produce similar scores. Internal consistency is usually measured with Coefficient Alpha, introduced by Cronbach in 1951, which is a statistic calculated from the pairwise correlations between items. Cronbach's alpha is a measure of the homogeneity of a test and its items. When the various items of a test are measuring the same construct (e.g. reading comprehension), then scores on the test items tend to covary. That is, test-takers will tend to answer in a similar manner across related items. The sections on a test that have adequate internal consistency will have items that are highly inter-correlated. Cronbach's alpha is a number between negative infinity and one, and as a rule of thumb, any coefficient alpha above 0.7 indicates acceptable reliability (Crocker & Algina, 1986).

When it comes to reliability, TOEFL iBT research has a lot to say. Reliability can be increased by analyzing the psychometric properties of a test, e.g. difficulty and discrimination indices of test tasks and internal consistency of test tasks. For the objectively scored sections of the TOEFL, internal consistency is the extent to which tasks measuring a language construct correlate with one another. In case of open-ended speaking and writing sections of the test, consistency is measured in terms of rater agreement (inter-rater reliability) (Chalhoub-Deville & Turner, 2000). "The hallmark of TOEFL… is its psychometric qualities with a strong emphasis on reliability. ETS adheres to a more psychometric approach to test construction" (Chalhoub-Deville & Turner, 2000, p. 524). According to Pierce (1994) and Spolsky (1995), the processes by which TOEFL test items are developed are guided largely by psychometric research. Furthermore, the selected-response task type used in the reading and speaking sections of the test help increase reliability and decrease the measurement error. ETS has published several reports on the reliability coefficients of the four different sections of the TOEFL, and the alpha coefficients measured by ETS researchers are considered high, being larger than .70 (see an example in Table 4.1).

Table 2.1. TOEFL Reliability Estimates. Adapted from Educational Testing Service. (2011). Reliability and comparability of TOEFL iBT scores. *TOEFL iBT Research Insight, 1*(3), 1-8.

Score	Scale	Reliability Estimate	SEM
Reading	0-30	0.87	2.34
Listening	0-30	0.87	2.38
Speaking	0-30	0.86	1.57
Writing	0-30	0.80	2.14
Total	0-120	0.95	4.26

There are three versions of the TEOFL test: paper-based (TOEFL PBT), computer-based (TOEFL CBT), and internet-based (TOEFL iBT). TOEFL iBT has completely replaced CBT and is the most common type of the test. Despite the fact that many universities continue to publish the minimum entry requirements for TOEFL PBT, the iBT is the version taken by almost all college applicants who submit TOEFL scores. TOEFL iBT has four different sections: reading, listening, speaking, writing. While the items in the reading and listening sections are exclusively in multiple-choice format, the items in the speaking and writing sections require extended responses. Some of the prompts in these two sections are integrated and require students to use in

combination several different language skills: listening, reading and speaking. The independent speaking and writing items on the iBT are all open-ended.

As a test highly based on research with multiple reports published about its reliability and validity, the TOEFL test is extensively used for university admissions in the U.S. The internet-based version of the test (TOEFL iBT) was introduced in the United States in September 2005. TOEFL iBT was developed in response to a need for a test battery that would be used for university admissions to measure prospective students' ability to communicate in English in academic settings (Alderson, 2009; ETS, 2008; Sawaki, Stricker, & Oranje, 2009).

The validity of the TOEFL

The purpose of TOEFL is to measure the English proficiency of non-native speakers with regard to academic work and the intent to study in institutions of higher education in North America. The TOEFL was developed with the emergence of a demand for international scientific and technological communication in 1945 after World War II. The need to assess the English proficiency levels of individuals appeared as a result of an increasing demand among university students for effective communication in international academic settings. The growth of English as the language of academia has had a major impact on the widespread use of the TOEFL test all around the world. In 1962, the representatives of a number of governmental and private organizations formed a national council to address the issue of English language proficiency for non-native individuals who wished to study at universities in North America. The council came to the conclusion that a thorough English proficiency test should be constructed and administered to gauge the English proficiency of non-native speakers who wish to study or work in the U.S. The first TOEFL test, launched in 1964, was highly objective and prioritized reliability over validity (Elder, 2017; Spolsky, 1995). However, pressure to keep up with the research in the field of language assessment on the importance of direct and semi-direct assessment of language skills prompted ETS to make modifications to the test.

ETS claims that the TOEFL iBT measures a test taker's ability to combine listening, reading, speaking, and writing skills (via integrated test items) to engage in academic work and reports that the iBT is used by more than 11,000 academic institutions all around the word (ets.com). The TOEFL has gone through several major revisions since its initial design in 1964. These revisions were motivated by research on the nature of language proficiency and advances

in technology. Extensive research on all the aspects of the test and appeal to Target Language Use (TLU) situation by ETS researchers has been one of the main features of the test, making it face valid and popular among test score users.

In their ETS Research Report, Enright & Tyson (2011) make several propositions in support of the validity of the TOEFL, and then, present evidence for their arguments. Their first proposition "is that the test content is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings" (p.6). At least Three studies (i.e. Biber et al., 2004; Cumming, Grant, Mulcahy-Ernt, & Powers, 2005; Rosenfeld, Leung, & Oltman, 2001) provide evidence for this proposition in the TOEFL validity argument. Biber et al. (2004) analyzed a corpus of 1.67 million words of spoken language in academic contexts. The linguistic features they found in the corpus were then used to develop items for the TOEFL lectures and conversations. The main purpose of this research was to establish authenticity and representativeness for the TOEFL iBT test. Rosenfeld et al. (2001) administered a survey to undergraduate and graduate faculty and students to evaluate the importance of a variety of academic tasks to be used in the TOEFL test. Cumming et al. (2005) interviewed ESL teachers and sought their perception of the integrated speaking and writing tasks in the TOEFL iBT. The interviewees' perceptions about what these tasks represented and their suggestions about how they could be improved were used to further refine the integrated task characteristics.

According to the more recent definitions of test validity, it is important to validate a test based on evidence for each specific interpretation of a test in a specific context. Based on Toulmin's model of argument structure, the evidence-based design for validation has been discussed by many researchers, including Mislevy, Almond, & Lukas (2003). Toulmin's model has six categories: the *claim* which is the assertion that the argument-maker would like to prove, the *evidence* which is the grounds for supporting the claim, the *warrant(s)* or assumptions that link the grounds to the claim, and the *backing* which provides evidence for warrants, or *rebuttal(s)* to the claims. Mislevy et al. (2003) discuss the Evidence-centered assessment design (ECD) as an approach to constructing tests based on evidentiary argument and provide models that have been developed to implement the approach. They argue:

> Designing assessment products in such a framework ensures that the way in which evidence is gathered and interpreted is consistent with the underlying knowledge and purposes the assessment is intended to address. The common

> design architecture further ensures coordination among the work of different specialists, such as statisticians, task authors, delivery-process developers, and interface designers. While the primary focus of measurement specialists is building, fitting, testing, and reasoning from statistical models, this primer places such models into the context of the assessment process. (p. i)

Kane's (2013) argument-based approach to validity, which is based on Toulmin's (1958/2003) argument structure and Mislevy's work on ECD, argues that validation of a test requires the development of an argument supportive of target interpretations and uses. The if test interpretation and use are not well-defined, articulated, and defended, the argument(s) for interpretation and use will be weak. Kane's interpretation/use argument (IUA) states that "IUA includes all of the claims based on the test scores… some IUAs may focus on a particular use, while others may involve an interpretation in terms of a skill or disposition" (p. 2). Kane also mentions that no matter how detailed or broad an interpretation is, it must be supported by evidence, and its assumptions must be plausible. He calls his framework an "evidence-required" framework and mentions that the more ambitious a claim is, the more evidence is needed to justify the claim. The claim is one of the most ambitious when university staff are making decisions about the entry-level academic success of students based on TOEFL scores, and evidence is necessary to support that the assumptions behind those decisions are based on the knowledge necessary to make them.

Kane (2013) criticizes criterion-based approaches to test validity which were "gold standards for validity" between 1920 and 1950 by stating that "coming up with a suitable criterion can be difficult or impossible in many cases" (p. 5). Defining a specific criterion for test scores can be a simplistic approach for validation because in every kind of testing situation, we certainly want to measure a specific behavior that can manifest itself in various configurations. Simplifying a behavior to a specific criterion (e.g. GPA) and validating the test based on that criterion can be limiting and can lead to unfair consequences. Kane's proposed IUA can be specified in terms of a network of arguments leading from a test taker's observed performances to score uses and decisions after the test. The validity of interpretations and uses can then be determined in terms of the coherence of the network and the "plausibility of the assumptions". Kane makes three points about tailoring validation to proposed interpretations. The first one is that "test scores can have multiple possible interpretations/uses, and it is the proposed interpretation/use that is validated, not the test itself or the test scores" (p. 21). Kane states that "reasoning" for any interpretation and use of a test is the ultimate goal of IUA, and that is why

research studies focused on the use of language proficiency test scores by non-practitioners are inevitable in the "reasoning" and eventually the validation of the specific use of the test.

Chapelle, Enright, & Jamieson (2010) is the most complete application of Kane's validity argument approach. First, the approach emphasizes the development of an interpretive argument instead of the definition of a clear construct, which has always been difficult in language assessment. What forms the basis for the interpretation of scores in a specific context is the interpretive argument rather than construct definition. Second, validation research is a systematic and ongoing process of making inferences in the interpretive argument rather than making a list of potential validity evidence and validity threats. In addition, the interpretive argument is presented in a way that it is clear how other researchers can question, further investigate, or refute the validity argument. The ultimate goal of this approach to validity argument is to reach a conclusion about the adequacy of test score interpretation and use, and Kane's approach provides the validation researcher guidance necessary to embark on the validation process.

Central to Kane's validity argument is the 'target domain' which is the context of interest in which the language skill tested would be observed (Figure 2.1). Messick (1989) described the relations among the aspects of the schemata as content-related inferences, stating that "the key issues of content involve specification of the nature and boundaries of the domain as well as appraisal of the relevance and representativeness of the test items with respect to the domain" (Messick, 1989, p. 36). Both the specification of the domain and appraisal of relevance and representativeness are critical when it comes to the generalizability of a test's validity argument to various instances of test use. According to Chapelle (2011), "because of the importance of these connections between the domain and test tasks, in the TOEFL iBT interpretive argument, an inference, called domain description, is explicitly specified" (p. 22). Such inference 'links performance in the target domain to the observations of performance in the test domain' (Chapelle, Enright, & Jamieson, 2008, p. 14). In addition to validity research constantly conducted and published in various articles and research reports, a thorough validity argument for the intended interpretations and uses of TOEFL iBT is presented in a book by Chapelle, Enright, & Jamieson (2008). Motivated by task-based, TLU-focused validity research, integrated speaking and writing tasks were developed for TOEFL iBT in 2005 (ETS, 2011). These tasks "engage multiple skills to simulate language use in academic settings, and test materials that

reflect the reading, listening, speaking, and writing demands of real-world academic environments", which is the main focus of target domain validity argument (p. 2).

> Validity research for the TOEFL iBT test has been guided by an argument-based approach (Kane, 2001) that helps to lay out the different assumptions or claims explaining how the test is supposed to work to provide meaningful information about a test taker's academic communicative competence in English. It also establishes the types of evidence needed to support these claims. Initial validity evidence for the TOEFL iBT test is compiled according to the argument-based approach in the book edited by Chapelle, Jamieson, and Enright (2008) (ETS, 2011, p. 7).

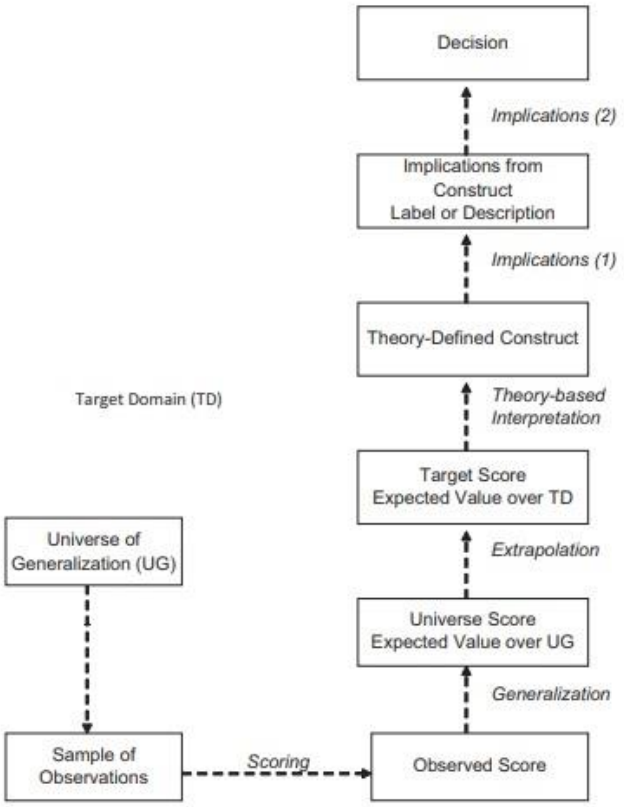


Figure 2.1. Elements and context of an interpretive argument. Adapted from Kane, 2006, p. 33

The use of TOEFL for ITA Verification

In American universities, assignment of international graduate students to teaching assistant positions has been a practice since the increase in the number of international students with the Open Doors policy after the 2000s (Elder, 2017; Ginther, 2003). Many international students in the Unites States are dependent on graduate teaching and research appointments to fund their studies "because federally funded grants and loans are restricted to citizens and

resident aliens" (Ginther, 2003, p. 58). Most graduate teaching assistantships require graduate students to engage in communication with undergraduate students, which makes adequate/efficient language use a necessity. In the 1980's, the difficulties undergraduate students had in comprehending and communicating with international teaching assistants (ITA) led to mandates being state legislatures being passed requiring universities to test ITAs for language proficiency before allowing them to hold teaching positions, especially in large state Research I universities (Ginther, 2003; Oppenheim, 1998; Thomas & Monoson, 1991). When it comes to ITA verification, speaking skills are usually the focus of assessment as the concern with ITAs' direct communication with undergraduates mostly involves the subskill of speaking.

There are direct and semi-direct methods of assessing speaking skills of ITAs. According to Ginther (2003), a speaking test is semi-direct when oral performance is captured across items that record test-takers' responses without the presence of a human interlocuter. In direct assessment of speaking ability, a human interlocuter is present to engage in conversation with the test-taker. Both generic tools, such as the TOEFL, and locally developed tools, such as Purdue's Oral English Proficiency Test (OEPT), are semi-direct methods of assessing ITAs. Direct methods usually involve Oral Proficiency Interview (OPI) tests, such as the American Council on the Teaching of Foreign Languages (ACTFL) OPI.

The TOEFL iBT speaking section is used in some academic institutions for ITA verification purposes "although the primary use of TOEFL iBT Speaking is to inform admission decisions regarding EFL/ESL applicants at English medium universities" (Xi, 2008, p.1). There are several research studies that focus on TOEFL's ability to assess the speaking skill for academic purposes (e.g., Butler, Eignor, Jones, McNamara, & Suomi, 2000; Cotos & Chung, 2018; Rosenfeld, Leung, & Oltman, 2001; Wylie & Tannenbaum, 2006; Wagner, 2016; Xi, 2008). Despite variation in how the different academic institutions respond to the need for ITA language proficiency assessment, the common aspect of the practice is an initial screening that requires prospective teaching assistants to demonstrate their oral proficiency by producing a language proficiency score at a certain level (Wagner, 2016).

Graduate applicants usually submit a language proficiency score as part of their graduate application, which is later implemented separately for ITA verification. The cut score required for admission is usually much lower than the cut score for ITA verification. For instance, Purdue Graduate School's cut score for the TOEFL speaking subskill is 18, whereas the TOEFL iBT cut

score for independent ITA certification is 27. ITA certification is usually a two-stage process. Students with very high scores on the speaking section of a language proficiency test submitted during the application process are usually certified for teaching. For those who enter with scores above the graduate school cutoff for speaking but below the speaking score requirement for teaching, or those without any speaking test scores, a local assessment of speaking skill is usually required. Although some scholars (e.g. Douglas, 1997; Powers & Powers, 2015) expressed their concern about the use of one isolated subskill score for assessment of English in academic contexts, the speaking score obtained from TOEFL iBT involves the subskill of listening and reading to some extent, as some of the speaking tasks on the TOEFL are integrated. However, Wagner (2016) argues that "in the ITA teaching domain, because teaching obviously involves both speaking and listening ability, it would seem advantageous to include both speaking and listening TOEFL iBT scores as predictors of teaching competence" (p. 3).

One of the most common tools for ITA certification is the now-retired Spoken Proficiency English Assessment Kit (SPEAK) developed by the ETS, which was administered on-site for ITA appointments after students arrived on campus. Since research found that SPEAK does not do very well in discriminating students in the middle range (Landa, 1988) and since ETS discontinued the SPEAK test, some universities started to use the TOEFL iBT Speaking score by itself but most use a locally-developed test for ITA certification. Universities typically use these tests to assign students into one of these three categories: pass, provisional pass, and fail. There are ITA training courses specifically designed for provisional passes. However, sometimes "the distributions of TOEFL speaking scores of the adjacent groups will overlap. Inevitably, those who are on the border between passing and provisionally passing and between provisionally passing and not passing are the toughest cases to classify" (Xi, 2008, p. 3). In addition, it is hard to distinguish among the various types of pedagogic discourse ITAs have for their job-related exchanges which would ensure that they can fulfill teaching tasks carried out in labs, recitations, or lectures (Axelson & Madden, 1994).

Xi (2008) sought to "provide criterion-related validity evidence for ITA screening decisions based on TOEFL Speaking scores and to evaluate the adequacy of using the scores for TA assignments" (p. 1). The researcher used two types of criterion measures in the study: 1. locally developed performance-based tests which were being used to select ITAs, and 2. ITA course instructors' recommendations. One of the major purposes of this study was to determine

the cut score at which the ITAs would be able to communicate effectively with undergraduate students. A secondary purpose was to find evidence for the overall effectiveness of the TOEFL speaking scores in ITA certification. The speaking section of the TOEFL test was delivered to participants in four different universities, and the optimal cut score was derived for each university separately using the receiver operating characteristic curve (ROC curve). The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold levels, which then helps to identify a cutoff point on the TOEFL speaking section that keeps false positives low. "The findings support the use of the TOEFL speaking test for ITA screening because TOEFL Speaking scores were reasonably correlated with scores on the local ITA-screening measures" (p. 41). The researcher also made TOEFL speaking cut score recommendations for pass, provisional pass, and fail, which were different for each institution included in the study. In a recent study, Cotos & Chung (2018) sought to "validate a secondary use of the TOEFL iBT Speaking scores for the purpose of certification of ITAs in English-medium universities" (p. 4). The researchers used Chapelle et al.'s (2008) TOEFL interpretive argument to conduct a domain analysis using the systemic functional linguistics (SFL) theory which "treats language as social semiotics, as a resource used for communication" (Halliday, 1978, as cited in Cotos & Chung, 2018). The researchers used the framework to identify the discourse units in two different corpora, i.e. a corpus of ITA discourse and a corpus of discourse in TOEFL iBT speaking responses, to investigate the hypothesis that "the language functions elicited by TOEFL iBT Speaking tasks are identified in authentic ITA discourse" (p.10). The results of the study indicated that TOEFL iBT Speaking tasks are able to test most of the language functions that exist in ITA discourse, suggesting that this test is valid as an ITA certification test since it accounts for the language abilities necessary to function effectively in the target language use domain.

In another validation study, Wagner (2016) investigated the effectiveness of using TOEFL iBT speaking and listening scores for ITA certification at an urban research university. The researcher correlated the TOEFL iBT test scores obtained from a group of students with the scores obtained from a local ITA screening test and undergraduate student evaluations of ITAs' use of language in teaching. One of the important aspects of this research was the use of TOEFL listening scores in combination with speaking scores since the researcher asserted that "listening comprehension is an important aspect of instructional language competence" (p. 1). The

researcher reported that there was no correlation between the end-of-semester evaluation of teaching effectiveness of the ITAs and TOEFL listening and speaking scores, which could indicate that the end-of-semester evaluations are not accurate measures of teaching effectiveness. However, the researcher reports that "Whereas TOEFL iBT Speaking scores predicted only a negligible percentage of the teaching competence scores, TOEFL iBT Listening scores accounted for more than 20% of the observers' assessment of ITAs' teaching competence" (p. 36). This finding provides empirical evidence for the importance of the use of listening scores in addition to speaking scores for ITA screening purposes, since currently, almost all ITA screening programs in universities use only the speaking score for initial screening. However, one of the major issues with Wagner's conclusion is that his participants were students who were already assigned the role of teaching assistants, which indicates that the participants were certified to teach either through some sort of assessment of language skills (i.e. score on a local or a standardized test) or through the completion of one or several ESL courses.

Many institutions use TOEFL iBT at some point during their ITA screening process. Some use the TOEFL as an initial screening to make decisions about prospective ITAs who need to take an on-site screening test, and some use it to make final decisions about TA work assignments. Despite arguments for the use of TOEFL iBT speaking score for ITA verification, most institutions use a locally developed test that corresponds to their needs more closely. One reason for this is that the purpose of local tests extends beyond decision making about ITA appointments based on test results. Local tests are often embedded in a language program, and unlike standardized large-scale language tests, the language performances obtained from a local test can be used for diagnostic and research purposes (Dimova, Yan, & Ginther, 2020).

The Relationship Between TOEFL and CEFR

The Common European Framework of Reference (CEFR) has become the most influential international standard for describing second language proficiency. The CEFR was developed in 1971 by the Council of Europe (Council of Europe, 2001). According to the Council of Europe, the CEFR provides "a shared basis for reflection and communication among the different partners in the field" (para. 3). The CEFR is a flexible tool that can be adapted to any language use and assessment context as a reference of language ability. The six-point scale begins at level A1 for beginners and continues to level C2, near-native language proficiency. Table 2 displays

the six-point global scale and the descriptors for each level of language proficiency. CEFR publishes more detailed descriptors (Council of Europe, 2018) for each of the four main language skills (i.e. Reading, Listening, Speaking, and Writing) and micro-skills within each macroskill (e.g., within speaking: describing experience, giving information, putting a case, public announcements addressing audiences, etc.).

Table 2.2. Common Reference levels in CEFR. *Adapted from https://www.coe.int/*

	Level	Descriptor
Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Admissions decisions based on language proficiency scores require applicants to meet a criterion, and a test score by itself is not meaningful if it does not clearly indicate what criterion the test-taker meets by getting the score. One method ETS uses to define its score levels to

facilitate the criterion-related decisions based on the TOEFL is by linking the scores on the TOEFL to the six-point scale on the CEFR, which is called mapping. Mapping is conducted through standard setting, which is usually based on the combination of expert informant judgements and the test data. It must be mentioned that CEFR is a global scale developed to be used as a reference document in various contexts, not for academic uses of languages only (Milanovic & Weir, 2010). Tannenbaum & Wylie (2008) conducted an analysis to map TOEFL iBT score to CEFR levels; their analyses were criticized for being too rigorous, resulting in higher test scores than necessary to reflect the English skills described for each CEFR level (Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015). Therefore, a revised version of the mapping for the relationship between TOEFL iBT scores and the CEFR was published by Papageorgiou et al. (2015). The researchers acknowledge in the article that because university admissions decision-makers' feedback was incorporated into the mapping, "the reasonableness of these revised cut scores and their impact on admissions needed to be investigated" (Papageorgiou et al., 2015, p. 3). The process of standard-setting has been criticized due to its inherent subjectivity and the fact that the final product can be different based on what method is used for standard-setting (North, 2014a).

Understanding the CEFR is beneficial in the context of graduate admissions because it equals understanding what level of language ability TOEFL scores correspond to, and why the B2 level on the CEFR is critical for getting admitted to a graduate program. A B2 level of English will allow graduate students to function in their academic context since at this level they are able to comprehend abstract, complex ideas and communicate with a sufficient degree of fluency and spontaneity. Papageorgiou et al. (2015) investigated whether their revised mapping and the cut scores recommended based on the revised mapping are reasonable. They reviewed the admissions web pages of 155 universities in Australia, Canada, the U.K., and the U.S. The researchers were specifically interested in the B2 level on the CEFR which is recommended as minimum language ability necessary to successfully engage in academic tasks. The revised subskill proficiency levels they proposed were much lower than Tannenbaum & Wylie's (2008) original cutoffs (Table 2.3). Papageorgiou et al. (2015) argue that lowering the cutoffs in relation to CEFR might lead to an increase of false positives in the admissions, meaning that there might be more students who struggle with the linguistic demands of their academic programs after they are admitted. However, As stated by Bridgeman, Cho, & DiPietro (2016):

> English language skills are a necessary but not sufficient condition for success in academic study for international students at a university in which English is the only or dominant language of instruction. Because of this necessary but not sufficient relationship, scores on a test of English language abilities for international students should show some relationship to initial success at a university in the United States, but that relationship should not be expected to be a strong one because of all the other factors that can influence initial success. These other factors could include quantitative skills, knowledge in specific content domains, and a host of non-cognitive attributes such as motivation, persistence, and grit (p. 308).

Table 2.3. TOEFL iBT Test Score Requirements in Relation to the Mapping of TOEFL iBT Test Scores at the B2 Level of the CEFR. *Adapted from Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels. Ed. James Carlson. New Jersey: Educational Testing Service.*

Country	N	TOEFL iBT total score range 87−109 (original B2 cut scores)	TOEFL iBT total score range 72−94 (revised B2 cut scores)
Australia	7	1	7
Canada	14	6	14
UK	13	6	10
US	83	30	52

Lowering the initial cutoffs will allow the admissions decision makers to examine a broader range of proficiency and consider the whole application package which will contribute to students' academic success. However, it must be ensured that the language proficiency of applicants is being taken into consideration during the subsequent stages of application review carried out by the decision-making faculty after applications leave the Graduate School. Because background knowledge is an important contributing factor to academic success, TOEFL iBT by itself can have some, but not an exclusive relationship to student success. Local tests designed to pick on specific areas of language proficiency are better predictors of student's academic language ability than the more generally academic tests such as the TOEFL iBT because the range of language ability assessed by a local test is more focused than the range assessed by a standardized test such as the TOEFL.

OEPT: Purdue's Local ITA Screening Test

Purdue's Oral English Proficiency Program (OEPP) was established in 1987 to screen for the language proficiency of prospective international teaching assistants. A growth in enrollment of international students in 1980s due to an increase in the interest of international students in

Research I universities in the United States led to universities' reliance on these students for research and teaching tasks (Haan, 2009). Later, with the assignment of teaching and teaching assistantship jobs to international students, the "foreign TA problem" took shape (Bailey, 1984; Ginther, 2003). According to Ginther (2003), in the 1980s, "the strength of the public perception of undergraduate difficulties with ITAs led to the establishment of mandates … requiring that the oral English proficiency of prospective ITAs be certified before they would be allowed to have direct contact with undergraduates" (p. 59).

At Purdue, the discussion for the creation of an ITA English proficiency support program began in the 1979 report on International Education Programs. Various departments at Purdue began sending requests to the English Department for the establishment of a program to provide English language support for ITAs. In 1981, a course was designed and introduced by an instructor in the English department. However, Purdue staff and faculty in the English department gave proposals for the creation of a new intensive English program, and finally, in 1988, the "Statement on Oral English Competency for Non-Native English Speakers Employed as Graduate Teaching Assistants/Instructors" was released. "According to this statement, all TAs must display adequate English proficiency before being placed in a position requiring contact with undergraduates" (Haan, 2009, p. 75). The policy required the departments at Purdue to detect ITAs who lack sufficient English communication skills and introduce them to the OEPP for support.

Purdue's OEPP was established in 1987 to fulfill the Purdue University requirement which states that any student whose first language is not English, and who is to be appointed as a teaching assistant, must demonstrate sufficient oral English proficiency before beginning their appointment and before they engage in direct communication with undergraduate students. Despite the best intentions of this Purdue policy, it was decided that its real-life implementation must exclude students who hold office hours/help sessions and students who perform their TA duties as graders. Policies like this are often modified in practice to best fit to the needs of the university.

The OEPP uses the Oral English Proficiency Test (OEPT), a locally developed, computer-delivered, oral English proficiency test, to certify students and exempt them from or place them into OEPP's ENGL 620 course, Classroom Communication for International Graduate Students. Currently, a five-point holistic rating scale is being used by the OEPP to rate OEPT test-takers

(see the complete scale in Appendix A). Scores on the OEPT range from 35 to 55 with 5-point increments. Students who score 50 and 55 are "certified" for English language proficiency and are not required to enroll in ENGL 620. Students who score 45 are "borderline" and "minimally adequate for classroom with support". Therefore, they are allowed to hold teaching positions with concurrent enrollment in ENGL 620 (Classroom Communication for International Graduate Students) to get the necessary support. Students scoring 40 have "limited" English speaking proficiency and must enroll in ENGL 620 and be certified by class before holding teaching appointments. Students scoring 35 are those whose "language resources or ability to communicate is restricted" and who will "likely to need more than one semester of support" (OEPT Holistic Scale, Appendix A). Students who score 35 on the OEPT are not placed in ENGL 620 because they are likely to have difficulty achieving certification after only a single semester of instruction.

International graduate students in the United States are highly dependent on graduate assistantships, which include tuition waivers, to be able to pursue their degrees. According to the Bureau of Labor Statistics, as of May 2018, there were more than 130,000 students holding graduate assistantship employment in the United States. With Purdue's international graduate and professional students representing 39.8% of all students at this level of study (Purdue's International Students and Scholar's annual report), many international graduate students have teaching assistantship appointments. Graduate admissions committees' use of language test scores for student selection determines the pool of students that the OEPP tests every year for ITA verification. The OEPP is able to provide English speaking proficiency support to students who score 40 and 45 on the OEPT, however, since the speaking proficiency level of students who score 35 is restricted, the OEPP is not any longer able to provide them with any form of support.

The OEPP has been involved in language proficiency standard setting through collaboration with the Graduate School. In 2016, an English Proficiency Task Force, consisting of the Graduate School and OEPP members, was created by the Graduate School to review the English proficiency requirements of graduate admission for international students. The task force recommended to the then Dean of the Graduate School that the minimum cut scores for PhD candidates must be increased to 80 for the TOEFL total score and 20 for each TOEFL subskill score. The Task force also recommended raising awareness about better use of English language

proficiency test scores in the admission process and making recommendations about student selection using language proficiency scores to each individual department.

This study's investigation of graduate selection committees' student selection dynamics will enable the OEPP to provide the admissions decision makers with information to help them make more informed decisions in terms of which language proficiency score profile is more likely to pass the OEPT, and which language proficiency profile cannot be supported by the OEPP. In line with this mission, the OEPP sends out 'score use recommendations for graduate admissions' to all Purdue departments to inform them about the ability of students in various TOEFL score levels to perform on the OEPT, and what TOEFL score levels are more likely to need further English support. Recommendations made by the OEPP for TOEFL score use in admissions are motivated by Figure 2.3 which displays the number and percentage of students in various TOEFL speaking and total score categories across the five OEPT scores. According to Figure 2.3, as the students' TOEFL total and speaking sub-scores increase, so do the green bars which represent those who pass the OEPT. The graph also displays the importance of paying close attention to the subskill score of speaking in addition to the total score. Despite the fact that TOEFL cut scores for graduate admissions set by the graduate school are Writing 18, Speaking 18, Listening 14, and Reading 19, the score interpretation recommendation memo sent by the OEPP to Purdue departments each year states:

> With respect to selection of students for assistantships who come in with a TOEFL iBT total score of at least 100 with a Speaking subscale score of 22, these students have about a 50% chance of passing the OEPT. Students who come in with a TOEFL iBT total score of 100 or higher with a Speaking subscale score of 24 or higher are more likely to pass the OEPT.

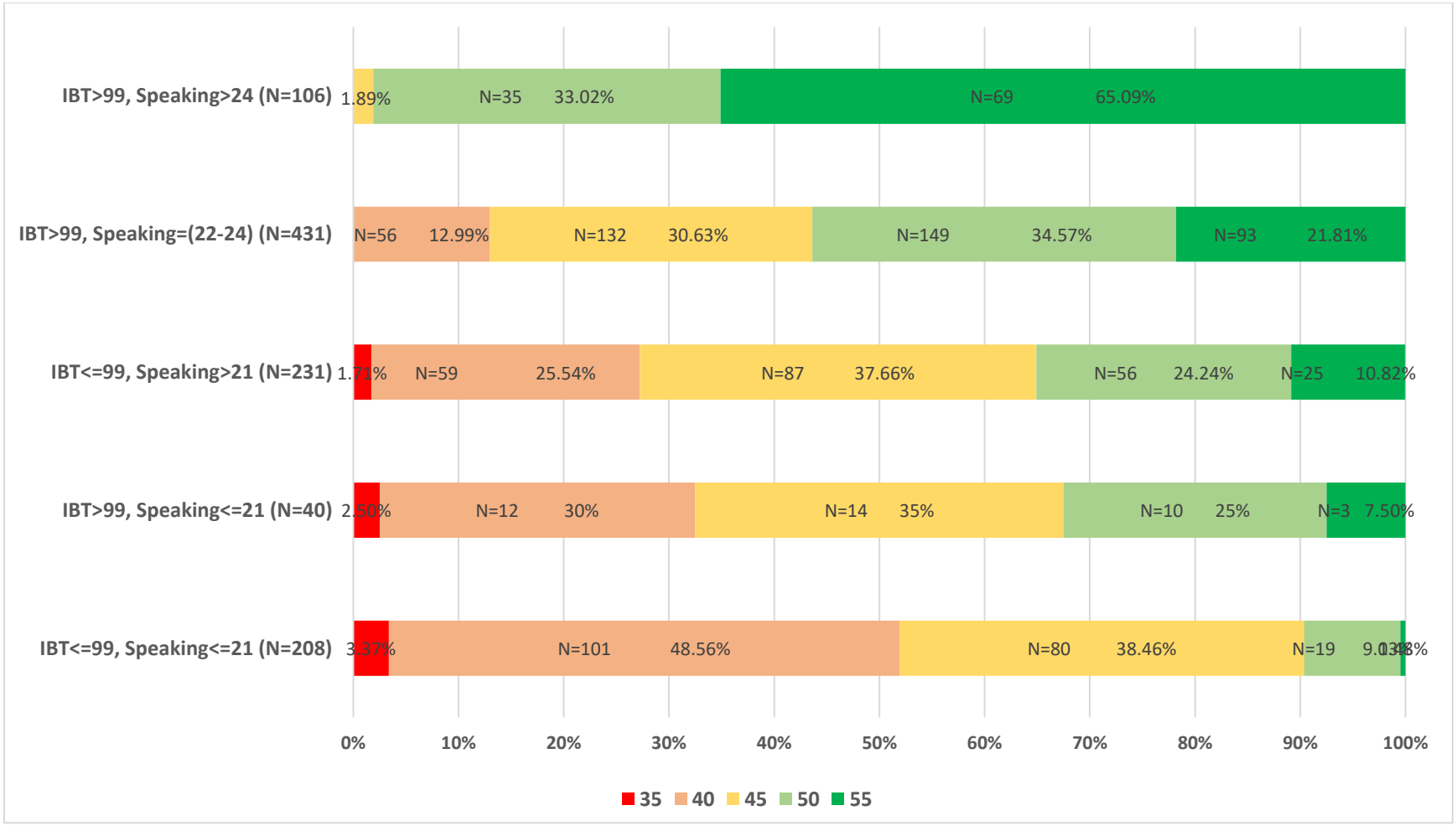
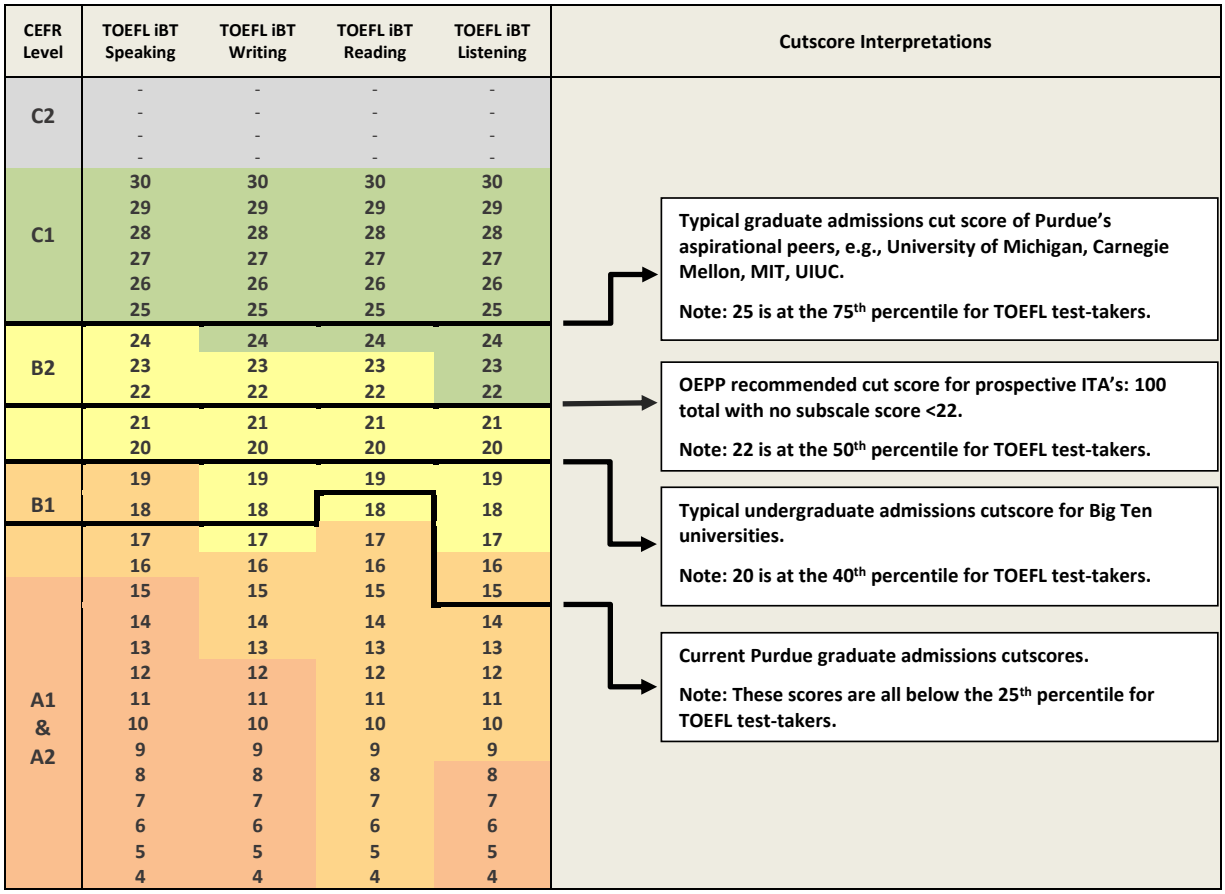


Figure 2.2. The Percentage of Students in Five different TOEFL Speaking Score Categories in Each OEPT Score Level (N=1016)

In addition to the annual score use memo, the OEPP distributes a TOEFL and CEFR mapping graph based on Papageorgiou et al.'s (2015) mapping results to all Purdue departments to inform them of the various CEFR levels that are represented by each TOEFL score on each subskill (Figure 2.4). As mentioned before, a minimum productive and receptive B2 level on the CEFR is recommended for successfully carrying out tasks related to graduate studies (refer to Table 2.2 for CEFR level descriptors). Figure 2.4 displays each CEFR level pertaining to each of the four subskills in five different colors. Levels A1 and A2 on the CEFR are displayed in the same color and denoted as insufficient level of English proficiency for academic activities. It is clear in the graph that Purdue graduate admissions cut scores lie in the border between B1 and B2 levels of language proficiency based on Papageorgiou et al. (2015); however, Papageorgiou et al.'s (2015) recommendations for TOEFL cut scores are liberal when compared to Tannenbaum & Wylie (2008). OEPP recommends selecting at the B2 level and above, stating that students who have at least an intermediate level of language proficiency can either be certified for teaching by the OEPT or benefit from further linguistic support.



A **C2** level of English is essentially the level expected of a first language speaker. C2 allows for reading and writing of any type on any subject, nuanced expression of emotions and opinions, and active participation in any academic or professional setting. TOEFL is not designed to reliably measure C2.

C1 is the level at which a student can comfortably participate in all graduate activities, including teaching.

B2 measures the level required to participate independently in higher level language interaction. It is typically the level required to be able to follow academic level instruction and to participate in academic education, including both coursework and student life. However, B2 is an advanced intermediate level of language proficiency. Students entering graduate studies at B2 benefit from language support, e.g., ENGL 620 or PLaCE short courses.

B1 is insufficient for full academic level participation in language activities. A student at this level could 'get by' in everyday situations independently. To be successful in communication in university settings, additional English language courses are required.

A1 and **A2** are insufficient levels for academic level participation.

Figure 2.3. TOEFL iBT/Common European Framework of Reference (CEFR) descriptors for participation in university activities. *Adapted from OEPP TOEFL Score Use Recommendation Memo*

2.2.2 Language Assessment Literacy (LAL) of Admissions Decision Makers

Graduate studies provide the opportunity for students to conduct research, gain specialized knowledge beyond what is learned at the undergraduate level, and develop a specialized skill set. A combination of factors contributes to the decision of admitting a graduate student into a graduate program: standardized test scores, research and work experience in the field, recommendation letters, writing samples, GPA, and the letter of intent. Research studies conducted to predict student success based on TOEFL and GRE scores (e.g. Kuncel & Hezlett, 2007; Kuncel et al., 2001; Sternberg & Williams, 1997) are motivated by admissions decision makers' concerns about the use of "explicit cutoffs or tacit minima" when considering the weight of standardized test scores in student admissions. Admissions committees are curious to know what effect excluding applicants based on their meeting a cut score has on the selection of students who nonetheless may be academically successful (Posselt, 2016).

The rise in the volume of graduate applications received by higher education institutions has led to the prevalence of deliberative bureaucracy in the process of admissions decision making. Due to the impossibility of involving a large number of faculty in openly discussing every application received by the graduate school, setting quantifiable standards can lead to speed, efficiency, and consistency in the admissions process at its initial stage. "Putting numbers to judgments in order to simplify comparisons among applicants" is now central to the graduate application review process (Posselt, 2016, p. 30). Since quantifying quality for graduate admissions is unavoidable due to the large number of applications received, it is important to investigate how cut scores are perceived and implemented by the graduate admissions committees after the applications leave the graduate school. Language proficiency, as one of the criteria in the selection of international graduate students, has often baffled the decision-making faculty.

Ginther & Elder (2014) investigated the process of admissions decision making in relation to language proficiency scores and report that there are varying viewpoints about language proficiency cut scores and their meaning to faculty. While most of their respondents were aware of the level of English at which their students were admitted, some others had misconceptions or uncertainties about what the university cut scores for English proficiency represent. The ultimate weight that standardized test scores carry in the process of admissions can be affected by these uncertainties through either overreliance or under-reliance on the cut scores set by the graduate

school. Therefore, addressing the misconceptions and uncertainties is crucial if informed admissions decisions made by the admissions committee members is the goal.

There is no doubt that testing and assessment of languages is gaining more weight in the world today. With the increase in global communication, research into the use of tests in facilitating these communications is also increasing. The growing literature on the assessment literacy of educators and language teachers reflects the importance of testing in today's educational settings. However, exploring the language assessment literacy literature quickly bares the truth about the extent to which a specific group of stakeholders was excluded from analyses; a very diverse group of individuals who are frequent users of the most renowned standardized language tests, such as the TOEFL, are university graduate admissions committees. The task of admitting students to various graduate programs is usually undertaken by the professors in various programs. However, there is little research about how this group of stakeholders can be better informed about the meaning of language test scores (Baker, 2016; O'Loughlin, 2013).

What is LAL?

Assessment literacy (AL) of educators has been a point of discussion in many research articles in the field of education. AL was traditionally defined as the basic knowledge/skills of assessment conventions that teachers and other test users need to have in order to justly measure student achievement (Xu & Brown, 2016). AL is an integral part of the teaching profession as teachers use assessment not only because they do small-scale classroom testing very frequently, but also because of the role assessment has in student learning (Black & William, 1998). No matter what teaching method is used, all teachers are involved in some assessment-related decision-making during their professions, and if teachers are not sufficiently informed about test development and use, assessments cannot be used appropriately (Stiggins, 2010). DeLuca, LaPointe-McEwan, & Luhanga (2016) "analyzed assessment literacy standards from five English-speaking countries (i.e., Australia, Canada, New Zealand, UK, and USA) plus mainland Europe to understand shifts in the assessment landscape over time and across regions" (p.251). Their results indicate that while recent (2010-present) assessment standards highlight the importance of formative assessment and Assessment for Learning, teachers' use of assessment is mostly compliant with older standards (1990–1999) which emphasized the selection and use of

summative and standardized assessment to make fair educational decisions. DeLuca et al. (2016) also maintain that we need to "establish the value and validity of assessment literacy instruments based on a close coupling with both assessment standards…and teachers' actual assessment practices (i.e., correspondence between what teachers say they do/know in assessment and how they actually assess in their classrooms)…" (p. 269).

Language assessment literacy is the knowledge that stakeholders involved in the assessment of language must have in order to reach sound decisions based on test scores. The discussion of LAL shifted in recent years to include all parties involved in the assessment process, and LAL is the impact of various practitioners' assessment knowledge on students' lives. The Standards for Teacher Competence in Educational Assessment of Students published by the American Federation of Teachers et al. (1990) specifies five domains in which LAL is meaningful:

1. Choosing and developing assessment methods
2. Administering, scoring, and interpreting assessment results
3. Using assessment results for decision-making
4. Communicating assessment results
5. Recognizing unethical, inappropriate assessment use and information

In our multilingual world today, the widespread assessment of English as a lingua franca has stimulated the need for assessment literacy among various stakeholders. As a result, there is an increasing need for language assessment specialists to consider more precisely what is meant by "language assessment literacy" and to clearly articulate what role it plays in the lives of language test-takers and their changing needs. Fulcher (2012) states that language teachers are now more than ever responsible for assessment due to several reasons; In the United States, after the No Child Left Behind Act, accountability gained more weight in education, both in local and global educational contexts. As mentioned in Malone's (2017) book chapter, in 2001, "passage of No Child Left Behind (NCLB) in the United States mandates annual assessment of the English language proficiency of all English language learners enrolled in elementary and secondary programs" (p. 226) The act prompted programs to start emphasizing the monitoring of student success in using English to learn content areas. Therefore, the need for LAL training among content and language educators is both politically rooted and socially felt.

The second reason behind the need for language tests and assessment literacy is the increase in international mobility and globalization. As stated in Roever & McNamara (2006), complex social roles are being played by language tests in the world today, no matter if the goal of the test is "economic competition" or "the unprecedented movement of peoples for reasons of education, economic advancement or sheer physical survival" (p. 243). Although teachers might not be directly involved in the development, administration and decision-making process of all kinds of language tests, they are affected by students' demands for teachers to "teach to test" and provide classes that are mainly assessment oriented (Fulcher, 2012). The third reason mentioned by Fulcher (2012) for the emergence of a need for LAL is the educationally beneficial aspect of tests. The feedback given by both the test itself and the teacher (based on the test results) is deemed valuable in many research studies.

Scarino (2013) states that there has been a tension between two competing paradigms in language assessment: "The tension is between traditional assessment, which tends to be aligned with cognitive views of learning and psychometric testing, and alternative assessment, which tends to be aligned with sociocultural views of learning" (p. 312). While formative assessment, classroom-based assessment, and performance assessment are considered more important than summative assessment in the second paradigm, many educators are less willing to implement them in their language classes according to DeLuca et al.'s (2016) study. Scarino states that the difference in these two assessment paradigms is philosophical and rooted in learning theories behind each paradigm. Therefore, the understanding of theories behind these two paradigms must be part of every LAL training program before teachers adopt the role of assessors. While the first paradigm might seem like a more direct and straightforward measure of student achievement, understanding the theories behind the second paradigm might shed more light on the costs and benefits of using it instead of the first paradigm and be deemed more valuable by teachers. Therefore "the understanding and appropriate use of assessment practices along with the knowledge of the theoretical and philosophical underpinnings" is important when it comes to LAL training (DeLuca & Klinger, 2010, as stated in Fulcher, 2012, p. 126). Although a comprehension of assessment paradigms is deemed important in LAL by several researchers, the kind of LAL needed in the context of university admissions is quite different since test score user in this context are non-practitioners in the field of language teaching and testing. The kind of

LAL necessary for the admissions committees and the decision-making faculty requires further research into their needs.

Policy Literacy, Admissions Literacy, and Language Assessment Literacy

There are three levels of score use in the context of university admissions, and transfer of information from one level to another is crucial. The first level consists of the language testers and test developers. The second level consists of policy makers who use information provided by the language testers to set policies which are used by the third and the most important level of score users in the context of university admissions: admissions decision makers and policy users. While policy makers must be responsible for the transfer the information from the first level to the third level, members of the first level, i.e. language assessment specialists, are also responsible for the transfer of adequate assessment-related information both to the second and the third levels of score use. These three groups of test users are like gears in the mechanism of admissions and the flow of information among these three groups is what will keep the mechanism running smoothly.

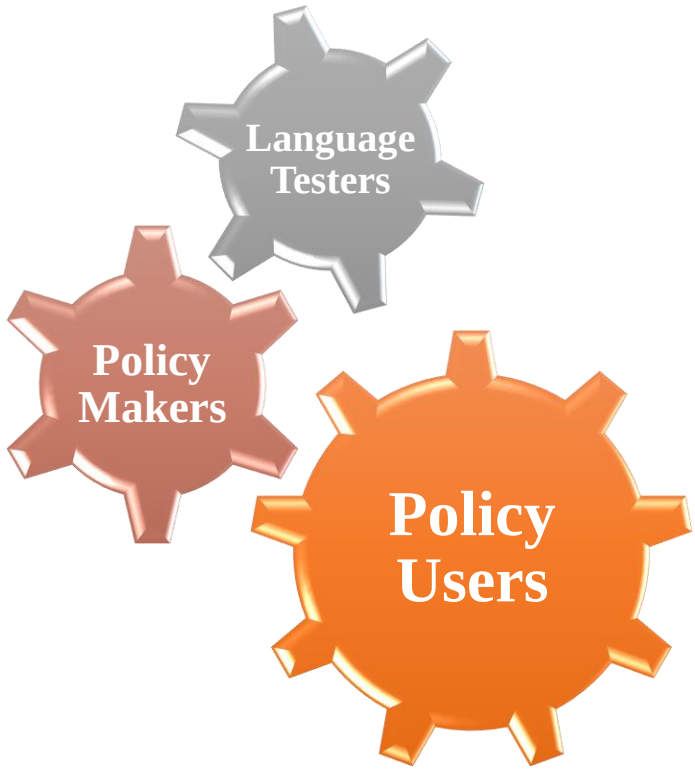


Figure 2.4. Levels of Language Proficiency Test Score Use in the Context of Graduate Admissions

As stated by Lo Bianco (2001), "policy processes … can evolve ambiguous relations with practitioners … Scholars can have an ambiguous relationship with policy since policy draws on research evidence in motivated ways that shape and frame this information for action" (p. 212). Policy literacy is a crucial part of language assessment literacy, because no matter how knowledgeable one is in the area of language assessment, the admissions policies governing selections in higher education institutions can outweigh the best practice possible in terms of language assessment. For instance, in a regular language assessment setting, the higher the applicants' language test scores are, the stronger they are considered to be. However, in the context of graduate admissions, the goal is to find a balance between language proficiency test scores and many other factors involved in the decision-making process (Figure 2.6). Therefore, the transfer of information between policy users and language assessment practitioners must be reciprocal, meaning that language assessment specialists must gain admissions literacy in order to be able to offer advice contributing to the general language proficiency literacy of the policy users. The kind of assessment which occurs in higher education institutions is a compromised one, based on Boltanski & Thevenot's (2006) theory of situated judgement, to maximize efficiency. Graduate admissions committees know that they need to select their next cohort of students with many factors in mind: student success, diversity, and all the other factor that would eventually contribute to the status of the university as a whole. In fact, the process of admissions involves the negotiation of hierarchies of priority; asserting best language assessment practices irrespective of the contextual hierarchy that exists in each admissions committee would be naïve. That said, it is not easy to identify and sort these hierarchies of priority since they vary from discipline to discipline, committee to committee, and even individual to individual. Therefore, when promoting language assessment practices in the context of admissions, it is important to analyze and document the practices of various departments and programs separately.

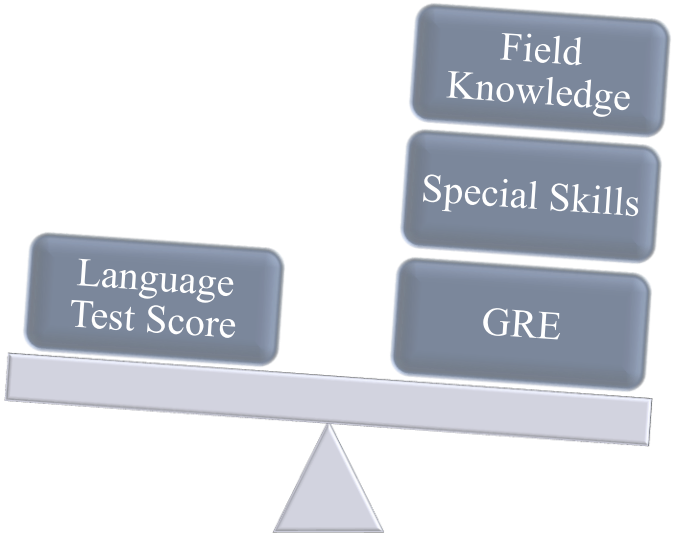


Figure 2.5. The Balance between Language Proficiency Test Scores and Other Factors Important in Graduate Admissions Decisions

LAL is a fairly new topic in the field of language testing and assessment and is still establishing itself. In order to establish a clear theoretical framework on which we can base our practice of increasing LAL among non-language practitioners who are language test users, more research is necessary.

Language Assessment Literacy or Language Proficiency Literacy?

In Case of LAL development of non-practitioners, graduate admissions committees in this case, one important consideration is the level of LAL they need to attain in order to make informed decisions based on test-takers' language proficiency scores. In other words, to what extent do non-practitioners need to be literate in Language Assessment? What are the areas of knowledge in the field of language assessment that will help them do their jobs? As mentioned earlier, admissions committees at universities are one of the main users of standardized language test results, but they are at the same time the group that can have misconceptions about language tests and the minimum cut scores (Ginther & Elder, 2014). Is it knowledge about language

assessment that will help them make more informed decisions on language scores or is it knowledge of language proficiency?

In the LAL literature, researchers have made various suggestions for LAL development of stakeholders. Boyles (2005) recommends the development of several competencies by stakeholders, including the ability to understand appropriate testing practices and interpreting the results of an assessment. Inbar-Lourie (2008) also suggested a framework of core competencies that includes "a body of knowledge" about appropriate language assessment. Davies (2008) outlines three elements for language assessment literacy: skills (how to do assessment), knowledge (what language assessment is), and principles (knowledge of validity and reliability issues). While all of these LAL development recommendations are relevant to the roles of teachers as testers, the knowledge admissions decision makers need to have about language assessment to make informed decisions about graduate applications does not have to be as comprehensive. It is therefore important to find the balance between what the decision-making faculty know or do not know about language test scores and what they need to know in order to perform their job duties, no more and no less. We must strive to keep these non-practitioners in their stretch zone by providing knowledge that is directly related to their needs and their identified skills gaps. What deciding faculty need to know must be directly related to what they want to know, what they perceive as important, and what is perceived as necessary by language assessment practitioners. In this case, it is hard to call our attempts Language Assessment Literacy because we do not want non-practitioners to gain knowledge, skills, and principles needed for assessment in general, but rather we want them to gain knowledge and skills necessary to use language proficiency test scores. Therefore, it might be more suitable to call our endeavors to increase non-practitioners' knowledge of language proficiency test scores 'Language Proficiency Literacy' (LPL) rather than Language Assessment Literacy, because what admissions decision-makers need to know in order to make informed decisions is not knowledge of language assessment, but rather knowledge of language proficiency test scores and admission cut scores represent.

LAL Research in the Context of Admissions

While there is a considerably large body of literature about the importance of LAL training among language teachers and language practitioners, the important role of LAL among another

group of stakeholders has been under-researched: university admissions committee members. In response to the growing number of people involved in the process of language assessment, a certain amount of LAL is required for admissions decision-makers to be able to make ethical and effective use of language proficiency measures. Recent scholarship has suggested that language assessment practitioners have not been playing the important role they could be playing in validating the use of language proficiency tests for admissions purposes through working with non-language practitioners with the goal of developing appropriate levels of LAL (Baker, Tsushima, & Wang, 2014; Taylor 2009).

O'Loughlin (2013) accuses the language testing community of paying more attention to reducing measurement error in tests that are being constructed rather than "trying to understand the risk of making decisions about the fate of human beings using fallible language tests" (pp. 364-365). His study tried to shed more light on this issue and investigated the assessment literacy needs of university staff with different roles in relation to the International English Language Testing System (IELTS) test. The study is wider in purpose since the researcher tried to analyze the LAL needs of university staff involved in various ways with the test. He recruited participants who were using the test in admission, marketing, and academic roles. He also distinguished between "subjective" needs (those identified by participants themselves) and "objective" needs (those identified by other parties such as the researcher himself). In the first phase of the study, the researcher used a survey with mostly multiple-choice items. In the second phase of the research, he used semi-structured interviews to have a more in-depth analysis of test-user needs.

O'Loughlin's (2013) results led to several noteworthy findings. Firstly, it was found that the participants needed IELTS literacy mostly for "advising prospective students about English language entry requirements" and making "student admission decisions" (p. 370). It is also noteworthy that very few test users stated that they needed IELTS literacy to set cut-off scores as university admission requirement. This indicates that "setting and revising minimum entry requirements were not frequently undertaken" (p. 370). Secondly, the interviews conducted in the next phase of the study revealed that the participant test users were only concerned with having access to "surface" information about the IELTS (i.e. the minimum score necessary for university admission). Eighty-four percent of the participants mentioned that they used only the university admission's webpage to access information about IELTS. These results suggest that

the inclusion of information about test-taker characteristics and the meaning of minimum cutoffs on universities' admissions website can be helpful in informing admissions decision makers about the use of language scores during admissions.

Similar to O'Loughlin's (2013), Baker et al. (2014) believe that language assessment community has not played the role it must play in educating test score users to develop the a level of LAL which would enable decision makers to make informed decisions based on English proficiency test scores, despite the fact that there are increasing numbers of international applicants each year. In their 2014 study, Baker et al. report the first phase of their three-phase LAL project in Canada. The research questions they intended to address were what level of LAL was needed for proficiency test score users involved in the process of university admissions and what materials could be useful in developing the necessary LAL. They sent out a needs-analysis survey to the contact lists of 53 institutions around Canada, and they received replies from 19 of these institutions. Fifty-eight percent of their respondents were admissions officers and 42% were admissions administrators. It is noteworthy that their survey contained more open-ended questions than O'Loughlin's (2013) and was less controlled.

Baker et al. (2014) report the results of the quantitative, close-ended section of their survey first. Unlike O'Loughlin (2013), 15 out of 19 institutions mentioned that they use proficiency test scores to compile admission information and "set policy". O'Loughlin (2013) reported that very few respondents mentioned involvement in policy making. Furthermore, when asked how flexible the minimum language policies were, eight out of 19 institutions mentioned that they would "never" accept students whose language proficiency scores are below the cutoff. Baker does not mention what the cutoffs of these institutions were and how the difference between various institutions' cutoffs can affect decision-makers' flexibility when using them.

Baker et al.'s (2014) qualitative results shed more light on what could be included in a LAL development project. They came up with four major themes after analyzing the open-ended questions in their survey. The first theme is related to the construct measured in language tests. Many respondents believed that a high score in a language test does not always translate into the actual ability to use the language, and that test scores are not always true reflections of students' language proficiency. The second theme was related to the predictive validity of language proficiency test and the confidence respondents had about whether higher language proficiency scores were predictors of success in university. The third theme was related to the role of the

university in the continued student language development. The respondents believed that the university is responsible for supporting students with regard to their language proficiency development even after they are admitted and enrolled. The last theme reflected respondents' need for more information about language proficiency tests.

In the second phase of the LAL development project, Baker (2016) investigated how LAL can be described for score users in admissions decisions and what materials are useful in developing an LAL base for admissions decision makers. To answer its research question, the study gathered data in the first phase of the research to get an overview of the LAL competency profiles of English test score users. The researchers then created materials to build the LAL base, and then analyzed decision makers' evaluation of the developed materials. They held workshops in eight different institutions across Canada and had 59 total workshop attendees. The workshop attendees came from various job classifications within the universities. Several types of data were gathered during and after the workshops to allow the analysis of the workshop's efficacy. Their results revealed that the workshop participants received the workshop materials well and thought they were directly relevant to some of their job-related responsibilities. The researcher mentions that the results of each workshop will be used in the transformation of the next workshop based on the feedback received from the attendees. These findings indicate that admissions decision makers are engaged and interested in being informed about the use of test-scores, and the provision of such information will likely be well-received by this group of test-users.

Ginther & Elder (2014) investigated the LAL of admissions decision-makers at Purdue University and the University of Melbourne to see how English proficiency test scores are used in the graduate admissions process at each institution, how familiar test score users are with the language proficiency tests commonly used in these universities for admissions, and the conceptions and misconceptions these test score users have about language testing done via various language proficiency instruments. They used a mixed-method design with survey and post-survey interviews to collect their data. The study has several eye-opening results about admissions officers' knowledge and use of TOEFL tests. While many of Ginther & Elder's (2014) respondents expressed confidence in the predictive validity of language proficiency tests, they differed in their perception of the cutoff scores set by the graduate school for admission to graduate programs; at Purdue, "52% of the respondents … indicated that they understood the

language proficiency requirements set by the university as minimal English-language proficiency requirements, while 38% of the respondents indicated that they believed the requirements represent adequate English-language proficiency" (p. 14). There were 3% of respondents who believed the cutoff scores represented an "advanced" level of language proficiency.

When the respondents were asked whether they considered the requirements "too low, appropriate, or too high", the majority of them believed that the requirements were either two low or appropriate, and only 1% at Purdue indicated that the requirements to be too high. When asked to rank the importance of total TOEFL scores and subscale scores in graduate student success at the university, "at Purdue, the percentage of respondents who ranked the components of language proficiency scores as very important or important ranged from a high of 86% for a total score to 79% for a listening subsection score" (p. 15).

When asked about their preferred testing method for English speaking proficiency, 98% of the respondents stated that they preferred oral interviews for speaking, and at least 96% favored human scoring for speaking. One of the most striking findings of Ginther & Elder (2014) was related to the survey question that asked respondents to indicate how familiar they are with TOEFL, IELTS, and PTE (Pearson Test of English). "At Purdue, the percentage reporting that they were not familiar with any version of the language proficiency tests listed ranged from 31% for the TOEFL PBT to 69% for IELTS and 75% for the PTE" (p. 17). Ginther & Elder's (2014) results indicate that the comparability of the scales across different tests is another issue that must be addressed. The paper-based TOEFL (TOEFL PBT) is still being administered in parts of Russia and universities still accept its scores. The scale used in TOEFL PBT is different from TOEFL iBT which is by far the most prevalently taken version of the test for admission in the U.S. IELTS is a standardized language proficiency test mostly used for admission into European, Australian, and Canadian universities, but many U.S. institutions, including Purdue, accept IELTS scores. Purdue no longer accepts the Pearson English Language Test (PTE), but due to COVID-19 pandemic, Purdue announced in Spring 2020 semester that it will temporarily accept Duolingo test scores for undergraduate admissions. The introduction of CEFR to decision makers as a reference which would enable them to compare across these scales can be useful in admissions LAL development practices. In sum, Ginther & Elder's (2014) results revealed that despite the widespread use of language proficiency scores for admission to higher education institutions and the general perception of the importance of English proficiency level for

international student success, admissions decision-makers had misconceptions about the meaning of cutoff scores and that "assessment literacy among the respondents was generally limited" (p. 26) The research also reports that decision makers "expressed interest in further information" about language test scores and the meaning of university cutoffs (p. 26). An important concept related to this discussion is consequential validity of a test. Messick's (1989) facets of validity framework highlights the importance of test score implications and social consequences of test use. "The consequential basis of test interpretation comprises the value implications of constructs and their associated measures" (Messick, 1987, p. 96). Bachman & Palmer (2010) refer to consequential validity as a major component of the unifying notion of validity and state that the consequences of a test must be one of the primary considerations in any validation process. Despite many efforts to provide validation evidence for the use of TOEFL iBT for making various interpretations, the consequential validity of TOEFL iBT score use by a large group of stakeholders, i.e. admissions committee, is under-investigated.

Linn (1998) discusses the issue of partitioning the responsibilities among the "actors" in the evaluation of assessment programs. He states that "although the lack of a professional accountability mechanism for policymakers does not lessen the responsibility of the body that sets a policy of test use, it does complicate the partitioning of responsibility for evaluating consequence" (p. 28). The accountability of the use of language tests by non-practitioners, who are defined as those who are not involved and do not have experience in the development and administration of language tests but use tests to make decisions about test-takers, has been overlooked by policymakers in higher education institutions. The assumptions this group of test users make about the nature of a test influences their decisions, and not always are these assumptions in line with the ones language test developers make. Therefore, finding and addressing the mismatches that exist between these two groups of test-users is beneficial (Pill & Harding, 2013).

Another mismatch in assumptions that could affect admissions decision making is between those of policy makers and policy users. Policy makers often provide the policy to the policy users without elaboration on how the policy was set and how it must be interpreted. Therefore, policy literacy development is an essential step in language assessment literacy development. Deygers & Malone (2019) emphasize the importance of research on policy makers' perspectives on language assessment and state that as much as it is important to ensure that policy makers are

on the same page with language assessment specialists and test makers, it is also important to ensure that policy users share the same assumptions with the policy makers. One of the most important findings of Ginther & Elder (2014) is directly related to the issue of policy setting and policy literacy on the part of policy users. They report that some faculty members expressed uncertainty about the meaning of minimum cut scores for language proficiency. Some participants even mentioned that they believed the cut scores meant "sufficient" proficiency rather than "minimum" proficiency. Awareness about characteristics of applicant groups, informed use of test scores, and the meaning of cutoffs can be raised by sending out memos, having workshops and information sessions, and adding more information to the admissions web page.

2.3 Conclusion

This chapter presented a review of the literature related to the use of language proficiency test scores in the process of graduate admissions. The chapter started with introducing a brief history of graduate admissions in the United States and the different models used in graduate admissions. The pros and cons of each graduate admission model was then discussed. In the following section of the chapter, a detailed review of the use of TOEFL in the graduate admissions process was presented and the validity of such use and the relationship between TOEFL scores and CEFR levels were discussed. The last section of the chapter presented a review of the role of language assessment literacy in the process of admissions, and what is in fact necessary in terms of assessment literacy for decision-makers to know in order to be able to make informed decisions when involved in graduate student selection.

CHAPTER 3. RESEARCH METHODS

The present research investigates the language assessment practices and test score use procedures of graduate admissions committees in various colleges and departments at Purdue University to reveal score use patterns when selecting graduate students using language proficiency as a selection criterion. This chapter presents an overview of the purpose of the study, the dataset used to obtain results, and the methods used for data analysis.

3.1 Purpose of Research

The present study analyzes the international graduate application dataset, which consists of Purdue graduate application information from Fall 2016 to Fall 2020, to examine the characteristics of graduate applicants with regard to their language proficiency in each college and large department at Purdue University. The study helps language testers gain admissions literacy before being able to offer Language Proficiency Literacy (LPL) to admissions decision-makers. The two important pieces of information researchers obtain from the analysis are who Purdue graduate applicants are in terms of language proficiency, and if there are patterns/trends in association with language proficiency scores when discipline is taken into account. The value of revealing the linguistic characteristics of applicants in various disciplines lies in the fact that the one-size-fits-all method is not ideal when it comes to LPL development. For instance, recommendations made for selecting students above a specific minimum level of language proficiency will not be taken into account if there are not enough applicants available above that minimum to choose from. International student diversity is another factor that can affect recommendations made about student selection. If a large department or college is seeking international diversity, selecting international students from a single country whose students are generally strong in English proficiency (e.g., India) may not be possible. In sum, the LPL recommendations made to the decision-making faculty need to be contextual and realistic, pertaining to each discipline's specific needs along with the available pool.

Before being able to offer LPL development opportunities to graduate admissions decision-makers, language testers need to gain admissions literacy in their specific academic context. One way this can be achieved is by analyzing graduate admissions within their specific

institutional contexts to see patterns of admission test score use in each discipline and what the linguistic characteristics of the rejected and admitted applicants are. Providing admissions decision makers with this information can be a very helpful step towards enhancing LPL in the context of graduate admissions. To make the optimal selections in the graduate admissions process, the decision-making faculty can benefit from finding out about the linguistic demographics of their department's applicants, the difference between the characteristics of their admitted and rejected applicants, and how they compare to other large departments within the university. Therefore, the present study directly benefits both the graduate decision-making faculty at Purdue and the researchers of the study whose ultimate goal is to enhance LPL in the context of graduate admissions. The research seeks to answer the following research questions:

1. What are the characteristics of Purdue's graduate applicant pool in terms of language proficiency test score distribution across admission and matriculation status?
2. How do the distributions of total and subskill TOEFL scores compare across the two major language backgrounds (Indian and Chinese) of admitted, rejected, and matriculated applicants?
3. What are the language proficiency profiles of admitted graduate applicants and is there an association between proficiency profile membership and applicants' language backgrounds?

3.1.1 The Need for Language Proficiency Literacy

One of the main aims of this research is to take one step towards supporting the graduate admissions committees make informed decisions when considering graduate students' language proficiency test scores. Ginther & Elder (2014) administered surveys and postsurvey interviews to admissions committees at Purdue University and the University of Melbourne to investigate "levels of knowledge about and uses of test scores in international graduate student admissions procedures" (p. 1). Ginther & Elder (2014) has several key findings that motivated the present study. The researchers asked survey respondents how influential they thought English-language proficiency test subskill and total scores were in making admissions decisions. The vast majority of Purdue University respondents indicated language proficiency as either very important or

important and indicated that their local programs' English proficiency standards are much higher than the Graduate School Standards. However, only 16% of the respondents reported being familiar with the TOEFL iBT, which is by far the most commonly used test for U.S. graduate admissions. Despite recognition of the importance of language proficiency, lack of confidence in English proficiency tests is also one of the findings in Ginther & Elder (2014), which can be argued in relation to a lack of familiarity with the language tests. One respondent indicated during the post-survey interviews that many students who meet the minimum requirement for language proficiency struggle once they are in their graduate program and also in the job market. "We just don't have enough information about proficiency tests at this point to make a statement about them but really need to educate ourselves better on the issue" (p. 19).

Ginther & Elder (2014) found that there was little use of English proficiency test scores beyond reliance on the minimum cutoffs set by the Graduate School and local programs. "Assessment literacy among the respondents was generally limited" report the researchers as 66% of their respondents at Purdue strongly disagreed or disagreed with the statement *I am knowledgeable about language testing and assessment*. Around 51% of the respondents reported that they are interested in receiving information about language testing and assessment to increase their assessment literacy.

At the Language Testing Research Colloquium (LTRC, 2011), Ginther & Elder presented their findings from a preliminary study conducted to investigate how the main users of TOEFL and IELTS, i.e. admissions decision makers, interpret English-language proficiency test scores obtained from these tests. One of the important discussions of the present study is that language assessment literacy needs to be defined based on the context in which selection practices are being employed. Ginther & Elder also emphasize that "different dimensions of assessment literacy may need to be prioritized" for different disciplines and within each academic domain. The present study also argues that LAL development is not quite relevant in the context of graduate admissions. What admissions decision makers need to familiarize themselves with is LPL which is divided by this researcher into two categories based on Pill & Harding's (2013) definition of LAL:

1. Understanding characteristics of language test-taker applicants
2. Understanding what language proficiency test scores and cut scores represent in the context of admissions

The present study will focus on the first category, i.e. understanding characteristics of language test-taker applicants, by analyzing Purdue's graduate admissions dataset and tabulating the findings in a clear manner. The study will focus on revealing language test score use patterns in each college and large department by analyzing the graduate admissions data.

3.1.2 Interview with the Associate Dean of Graduate School

The Associate Dean of Graduate School at Purdue, Dr. Thomas W. Atkinson, was interviewed by the researcher in regard to the history of English-language proficiency policy making and the current actions being taken to ensure the minimum cutoffs for English proficiency are functioning to benefit both the university and the decision-making faculty. Dr. Atkinson discussed the process of graduate admissions at Purdue and the role of the Graduate School in the student selection process. Purdue is a large public, R1, land-grant university, known for its strong Engineering and Agriculture programs. In 2017, Purdue had the fourth largest number of international students among U.S. public institutions and was eighth overall among more than 4,500 public and private institutions, according to a report issued by the Institute of International Education. With the large number of international students applying to graduate programs at Purdue, the use of English-language proficiency test scores gains more weight in the admissions process. Dr. Atkinson stated that graduate admissions at Purdue is decentralized, meaning that each departmental or program committee make their own selections based on their own set of values after the Graduate School sends them completed application files. However, there is some initial screening that affects the whole process because the Graduate School has set minimum cutoff scores for English-proficiency tests and will not consider for admission any applicant who has not met those cutoffs.

One of major themes of Ginther & Elder's (2011) study revealed Purdue faculty's widespread confidence in standardized language-proficiency tests despite a general lack of knowledge about the content of the test, test scores, and validity evidence. The participants, however, were aware of the importance of language proficiency test scores in making funding decisions. Both in Ginther & Elder (2011) and Ginther & Elder (2014), faculty expressed disinterest in gaining knowledge about "language proficiency testing" but expressed interest to learn about "English-language proficiency test scores" and willingness to "defer to expert judgment".

According to Dr. Atkinson, Purdue Graduate School TOEFL cutoffs were set in 2005 at 19 for reading, 14 for listening, and 18 for writing and speaking subskills, which add up to 69. However, Purdue Graduate Schools' cutoff for TOEFL total score is 77 which is slightly higher than the sum of subskill cutoff scores. These cutoffs were set by a committee who tried to align the minimum scores with those of peer institutions. During Ginther & Elder's (2011) interview with a Graduate School official, the interviewee stated:

> We are lower than other Big Ten universities, but I consider it important to follow the recommendations of the standard setting panel, and I would not want to hold back a bright, promising L2 speaker because of having low scores on a single measure.

In 2017, an English Proficiency Task Force, consisting of members from the OEPP and the Graduate School, was charged with the task of reviewing the English proficiency requirements set by the Graduate School. After several meetings and reviewing a great amount of research and admissions data, the Task Force made the following recommendations to the then Dean of the Graduate School, Dean Mark J.T. Smith:

1. Raise the minimum TOEFL iBT overall score to 80 with 20 required across the four subsections; recommend 88 with 22 across subscales for PhD degree applicants. Inform individual program departments that target yields are possible with higher selection.
2. Eliminate waivers to international degree-seeking applicants whose native language is not English but who have been conferred a baccalaureate, graduate, or professional degree within the last 24 months from an English-speaking institution in a country where English is the native language. Do not waive the requirement even if an applicant has obtained a degree from a US institution.
3. Raise awareness about better use of English language proficiency test scores in the selection/admission process. Emphasize that cutoff scores are minimums.
4. Provide English for academic purposes courses for all graduate students. These would be voluntary and equivalent to a two-semester sequence.
5. Establish the corresponding requirements for IELTS and Pearson, with subscales, which are equivalent to the TOEFL iBT requirements. (English Proficiency Task Force Report)

The recommendations made by the Task Force to increase the minimum TOEFL iBT scores were later rejected by the Graduate Council. While, according to Ginther & Elder (2014), only 3% of faculty believed that Graduate School's English-language requirements represent *advanced* English proficiency, there are 38% of Purdue faculty who believe that the minimum cutoffs represent *adequate* English proficiency. Therefore, the English Proficiency Task Force's third recommendation, i.e., raising awareness about better use of English language proficiency test scores in the selection/admission process and emphasizing that cutoff scores are minimums, is important. The OEPP sends out a TOEFL iBT score use and interpretation memo to graduate admissions committees every year with recommended cut scores for TOEFL iBT:

With respect to selection of students for assistantships who come in with a TOEFL iBT total score of at least 100 with a Speaking subscale score of 22, these students have about a 50% chance of passing the OEPT. Students who come in with a TOEFL iBT total score of 100 or higher with a Speaking subscale score of 24 or higher are more likely to pass the OEPT. (TOEFL iBT Score Use and Interpretation Memo, 2020, Appendix B.)

One purpose of this research is to make this memo more discipline-specific by including information about graduate applicants' English-language proficiency test scores for each college and large department. Creating discipline-specific test score use reports for each college and large department will help the decision-making faculty know who their English-language proficiency test-taker applicants are, and how much linguistic support they would need after they start their graduate programs.

3.2 Data Analysis

The data used for the analyses in this research was obtained in November 2019 from Purdue's Graduate School, Office of Information Management and Analysis. The dataset consisted of Purdue Graduate Applicants' admission data for the academic years of 2016/17, 2017/18, 2018/19, and the Fall semester of 2020. The raw data received from the Graduate School consisted of 70,925 entries and included information about applicants' date of application, academic college, academic department, admission status, matriculation status, graduate level (i.e., master's vs. doctoral), citizenship status (international vs. domestic), country of citizenship, native language, GRE scores, TOEFL total and subskill scores, and IELTS overall and subskill scores. Three sets of analyses were conducted for the purpose of increasing

admissions committees' awareness towards Purdue applicants' language proficiency: 1) tabulation and graphing of descriptive data in an easily comprehensible way, 2) generating TOEFL score distribution graphs, and 3) comparing groups of applicants using a Cluster Analysis procedure. The SAS software package and the SPSS software package were used for data cleaning and analysis, and Microsoft Excel was used for graphing.

3.2.1 Tabulation and Graphing of Descriptive Data

Tabulation and graphing of descriptive statistics involve arranging, summarizing, and presenting the raw data in such a way that useful and comprehensible information is produced. By tabulating a large number of data points for each college of major department, the research intended to facilitate comparison for the decision-making faculty and bring out essential features of the admissions data. This process meant to reduce the bulk of information in the raw data in a simplified and meaningful form so that it could be used by the admissions committees.

Graphing was used in the research for displaying admission and matriculation rates for international and domestic students in each academic college. For this analysis, the SAS software package was used for data cleaning and analysis, and Microsoft Excel was used for graphing the results obtained from SAS.

Pie charts and bar charts are used in the research to compare the number of international applicants to domestic applicants in order to consider the plausibility of the argument that some disciplines might not have enough numbers of international applicants to be able to make better selections when it comes to English proficiency of incoming international students. Pie charts are also used in the research to display the linguistic international diversity of different colleges and the percentage of admitted applicants in five TOEFL speaking subskill score categories, i.e., *>27*, *24-26*, *20-23*, *<20*, and *No Score*. The *No Score* category refers to applicants who did not provide the Graduate School with English proficiency test scores at the time of their application. Most of the students in this category are those who had studied in an English-speaking institution for at least 24 months. This Graduate School policy was later changed and the amount of time necessary to be exempt from submitting English proficiency scores was increased to 36 months.

3.2.2 Distribution of TOEFL Scores by Admission Status

A histogram is used to plot the frequency of observations for a specific variable in a continuous dataset that has been divided into classes. In this research, graduate applicants' TOEFL score data were used to compare the distribution of TOEFL speaking scores in five score categories (i.e., 0-17, 18-20, 21-23, 24-26, 27-30) across admission and matriculation status. This kind of graphing helps admissions committees in different colleges and large departments to see how many applicants in each TOEFL score category they have admitted and rejected, and how many of the admitted students in each TOEFL score category were later matriculated. This will help the decision-making faculty see how their department/college has considered English-language proficiency in the admissions process along with other qualifications in students' application files, and how they can improve in terms of admitting higher proficiency students that meet the other admissions requirements or match their specific research area. Ideally, a college's TOEFL score distribution would resemble Figure 3.1. where the distribution of admitted students is skewed to the right, indicating that more high proficiency students are admitted, and the distribution of rejected students is skewed to the left, indicating that most low proficiency students are rejected. The research will examine how the distribution of rejected applicants' TOEFL scores compares to those of the admitted applicants in each college/department. For this analysis, the SAS software package was used in data cleaning and analysis, and Microsoft Excel was used for graphing the results obtained from SAS.

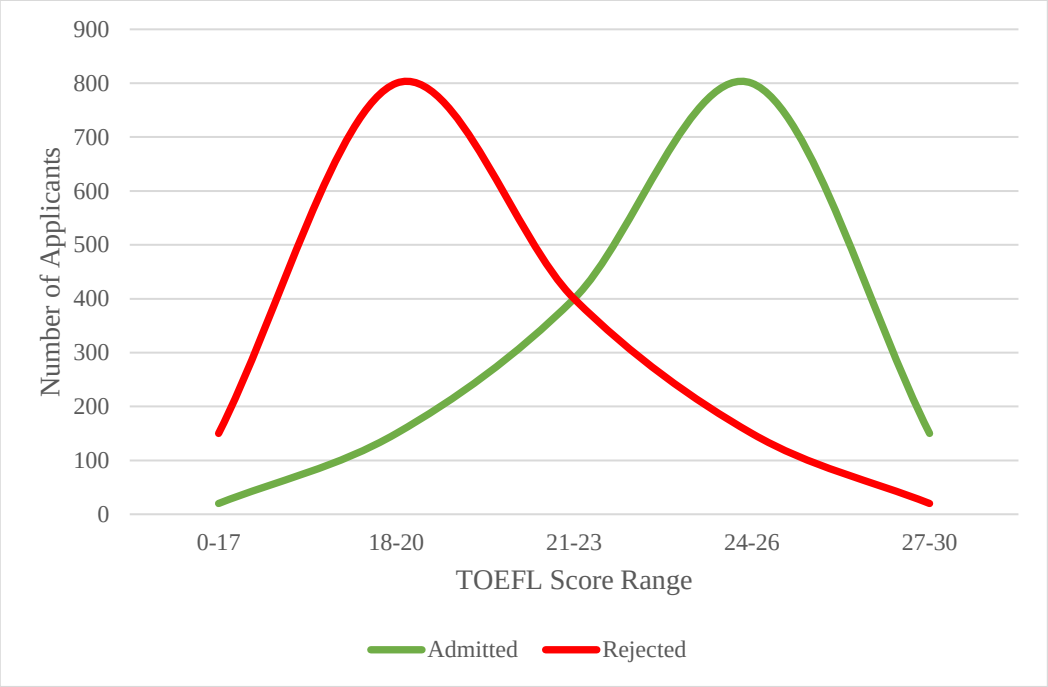


Figure 3.1. A Hypothetical TOEFL Subskill Score Distribution of Admitted and Rejected Applicants

3.2.3 The Cluster Analysis Procedure

The language proficiency profiles of applicants from the two different language backgrounds of Chinese and Indian are quite dissimilar. The language background groups that we included in the analysis were Chinese and Indian, since the majority of applicants belong to either group. A close analysis of the interaction that might exist between varying test score profiles of Purdue applicants and the admissions practices is important. A Hierarchical Cluster Analysis is conducted to study the English-language proficiency profiles of Purdue's graduate applicants (Ginther & Yan, 2018). The cluster analysis procedure, which is conducted using SPSS software package, is used to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that starts with each case in a separate cluster and combines clusters until only one is left. The study uses agglomeration coefficients to create scree plots that reveal the number of clusters we have in the language proficiency scores submitted by admitted applicants at the time of their application to Purdue. The scree plot will show the researcher how many distinguishable clusters there are in the data. The analysis will also show the centroids for each cluster in each group. The average distance of the scores from the cluster centroid is a measure of the variability of observations within each cluster. Generally, a cluster

that has a smaller average distance from the centroid is more compact than a cluster that has a larger average distance. Before conducting the analysis, we were expecting that the cluster centroids for English proficiency profiles of admitted students reveal three profiles: 1) the *unbalanced* English proficiency profile which consists of students who have higher scores across one or two subskills and lower scores on other subskills, for instance Reading, Listening, Writing > Speaking, or Reading, Listening > Speaking, writing 2) the *balanced medium* profile which consists of students who have moderate scores across all four subskills, and 3) the *balanced high* profile which represents applicants who have high scores across all four subskills. We do not expect to see a *balanced low* profile since the majority of admitted applicants have higher scores on at least one or two subskills.

The next step in the analyses would be to see if there is an association between language background and belonging to one of the linguistic profiles found in the cluster analysis. The cluster analysis results are used to conduct a Chi-square analysis to see if students' profile membership was related to their language background. In other words, the study sought to see if students from a specific language background were more likely to have *balanced* or *unbalanced* language profiles. The language background groups that we included in the analysis were Chinese and Indian, since the majority of applicants belong to either group. The association between language background and score profiles, and how students with different language profiles are being selected during the admissions process greatly affects our practices at the OEPP. In the unbalanced profiles, total TOEFL score is inflated due to high reading and listening scores. If admissions decision makers are selecting based on the total score rather than subskill scores, there will be a greater chance for the enrollment of students who do not have the minimum level of speaking proficiency to be able to perform their teaching assistantship job duties independently.

3.3 Conclusion

This chapter laid out the purpose of the study and the rationale behind the analyses conducted on Purdue's graduate admissions data. Research questions were presented and the types of analyses conducted were discussed in detail. The next chapter will present the results of the study and discuss the results in relation to the literature.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter consists of three major sections. In the first section, the characteristics of Purdue's international graduate applicants in selected Purdue colleges are compared and discussed in regard to the literature and the characteristics of international graduate applicants in the U.S. The next section summarizes the linguistic characteristics of Purdue's international graduate applicants in selected colleges and outlines applicants' linguistic profiles and their L1 backgrounds. The importance of communicating this information with graduate admissions committees and its contribution to language proficiency literacy development in the context of graduate admissions is also discussed. The last section of the chapter deals exclusively with the distribution of TOEFL subskill scores at Purdue's department of Engineering across three groups of students: admitted, rejected, and matriculated. The meaning and implications of the difference between the distribution of the speaking subskill and the other three subskills are discussed.

4.1 International Applicant Pool

The number of international students pursuing graduate education continues to grow world-wide. Getting a Ph.D. degree is now one of the most central prerequisites for faculty positions and promotions in many professional careers. The quality of graduate education in the Unites States attracts international graduate degree-seeking students from all around the world. In addition, economic and technological development in other parts of the world is one reason for increasing interest in graduate studies in the U.S. (Posselt, 2016). According to the National Center for Education Statistics (2019), "Between 2000 and 2017, total postbaccalaureate enrollment increased by 39% (from 2.2 million to 3.0 million students). By 2028, postbaccalaureate enrollment is projected to increase to 3.1 million students" (p. 1). According to the National Foundation for American Policy (2013), international student enrollment in the United States has contributed $24.7 billion to the U.S. economy. In 2017 and 2018, there was a slight decrease in the number of international graduate student enrollment in the U.S., a probable result of President Trump's Travel Ban and other governmental immigration policies. New enrollments fell 5.5% at the graduate level from 2016-17 to 2017-18. However, according to the Council of Graduate Schools (CGS), master's level applications increased by 1.4% and doctoral

level applications increased by 4.1% between Fall 2017 and Fall 2018. In addition, first-time enrollment in graduate programs grew by 2.0% for master's and 2.9% for doctoral degrees (Okahana & Zhou, 2019a). Despite recent declines in international graduate student enrollment (Okahana & Zhou, 2019a & 2019b), overall graduate enrollment at U.S. universities continues to grow. Since the workforce demands for graduate degree holders are also growing in the United States, such increase in graduate enrollments is not surprising. According to the U.S. Bureau of Labor Statistics (2019), jobs that require graduate degrees at the entry level are expected to increase by 13.7% for master's and 9.0% for Ph.D. between 2018 and 2028. The percentage of international students in various graduate programs varies by discipline. Nationwide, this percentage is as high as 81% for Electrical Engineering. The number and characteristics of graduate degree seeking applicants to various graduate programs in the U.S. graduate schools varies from discipline to discipline (National Foundation for American Policy, 2017).

The way the graduate admissions process is carried out is perceived as an indicator of how selective the educational institution is. However, "merit is always conditional" (Posselt, 2016, p.7). Where we draw the line between those few who are admitted and those who are considered inadmissible can vary greatly from discipline to discipline and program to program. As mentioned in chapter two, most graduate schools in North America use the 'holistic file review' method to review graduate applications. An inherent quality of the holistic method is its contribution to diversity by including each and every application in the review process. One single weakness in application materials should not exclude any applicant from being considered for admission. However, graduate schools usually set basic standards for admissions, which decreases the number of applications that are sent to departments and programs for review (Kent & McCarthy, 2016). The graduate school admission requirements in some universities are easy to meet for most graduate applicants while other schools choose to set higher standards for the initial graduate-school level review process. For instance, the minimum graduate school admission requirement for language proficiency at the University of Illinois at Urbana-Champaign (UIUC) is a total TOEFL score of 102, whereas the cut score of Purdue University graduate school is a total TOEFL score of 80 and subskill scores of 19 for reading, 14 for listening, 18 for speaking, and 18 for writing. The difference between these minimum proficiency requirements set by the Graduate School greatly affects the pool of applicants from which graduate programs admit applicants.

"Admission may officially be a matter of choosing students, but the selection process itself is an institutionalized compromise that balances and reflects multiple, sometimes competing, faculty values" states Posselt (2016, p. 36). Posselt found in her research that one of the most frequently debated student selection challenges was the extent to which faculty needed to impose the numerical thresholds for the TOEFL set by graduate schools or programs. The answer to this question is greatly dependent on 1) the nature of the graduate program, and 2) the characteristics of international applicants in each graduate program. According to National Foundation for American Policy (2013), in some graduate programs in the U.S., up to 70% of all graduate enrollment consists of international students (Table 4.1).

Table 4.1. Full-time Graduate Students and the Percent of International Students by Field (2010) – Adapted from National Science Foundation webcaspar.nsf.gov

Field	Percent of International Students	Number of Full-time Graduate Students – International Students	Number of Full-time Graduate Students – U.S. Students
Electrical Engineering	70.3%	21,073	8,904
Computer Science	63.2%	20,710	12,072
Industrial Engineering	60.4%	5,057	3,314
Economics	55.4%	7,587	6,117
Chemical Engineering	53.4%	4,012	3,504
Materials Engineering	52.1%	2,660	2,891
Mechanical Engineering	50.2%	8,352	8,273
Mathematics & Statistics	44.5%	7,840	9,766
Physics	43.7%	5,716	7,369
Civil Engineering	43.7%	6,202	7,989
Other Engineering	42.1%	7,279	9,992
Chemistry	40.3%	8,059	11,952

When comparing the pool of international applicants in various disciplines at Purdue University, the numbers are quite different by program. In STEM disciplines, the number of international applicants is much higher than that of non-STEM disciplines. For instance, at Purdue's College of Engineering, which offers M.S. and Ph.D. degrees in its various departments including Electrical & Computer Engineering, Mechanical Engineering, Civil Engineering, Aeronautics & Astronautics, and Industrial Engineering, 22% of applicants consist of domestic students, while 78% are international students. However, while 78% of all the applicants are international, only 59% of those are admitted. The percentage of matriculated applicants is 51%

international and 49% domestic (Figure 4.1), which indicates that admission is less competitive for domestic students. This trend is also true for other STEM colleges such as Science and Polytechnic. These numbers are in line with the general trend in engineering reported by the National Foundation for American Policy.

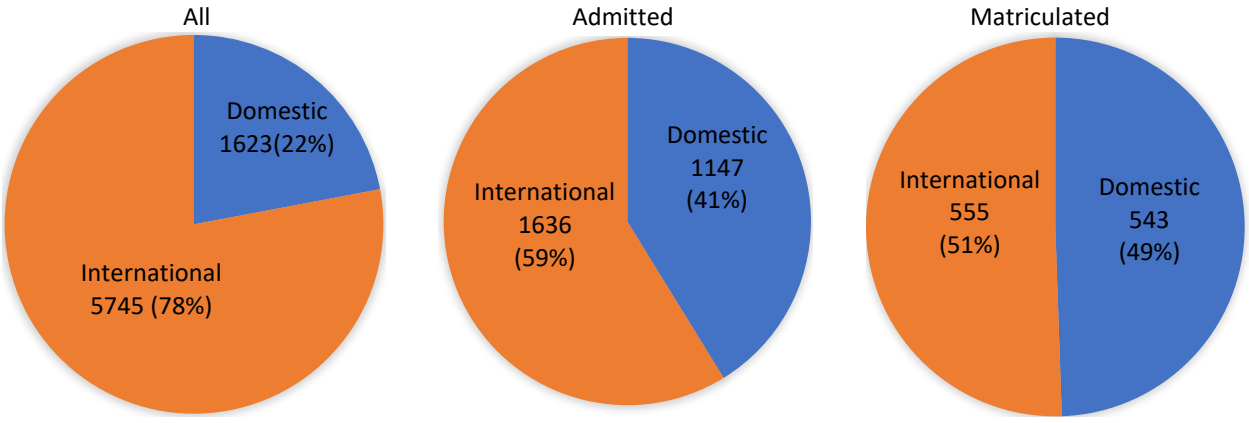


Figure 4.1. Number of All, Admitted, and Matriculated International and Domestic Applicants in College of Engineering – AY 2018/19

In non-STEM Purdue programs such as programs in the College of Liberal Arts or College of Education, the number of domestic applicants is much higher. Purdue's College of Liberal Arts offers M.S. and Ph.D. degrees in its various departments, the largest of which are the *Communication Department* and the *English Department*. Seventy-three percent of applicants to the college consist of domestic students, while 27% are international students. Fourteen percent of those international applicants are admitted, and the percentage of matriculated applicants is 9% international and 91% domestic (Figure 4.2). It is easier to find nation-wide enrollment information for Science and Engineering programs than for non-stem disciplines, which indicates that the nation-wide enrollment trend for non-stem programs is similar to Purdue's.

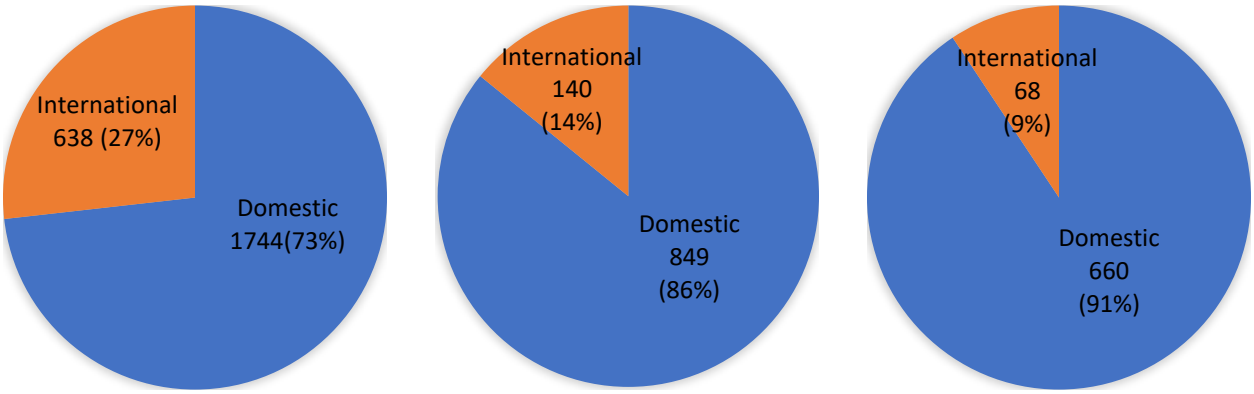


Figure 4.2. Number of International and Domestic Applicants by Admission Status in College of Liberal Arts – AY 2017/18, 18/19, FALL 2019

The three largest departments in the College of Engineering are *Electrical & Computer Engineering*, *Mechanical Engineering*, and *Civil Engineering*, accounting for 56% of all the applicants in this college. In all three large departments in the College of Engineering, the number of international applicants is higher than the number of domestic applicants. In *Electrical & Computer Engineering*, 24% of all international applicants are admitted, which indicates that the programs in this department are selective of and competitive for international students. However, in comparison, 59% of all domestic applicants in this department are admitted, indicating that getting into this college is more competitive for international applicants than domestic applicants. In the *Department of Mechanical Engineering*, 19% of international and 62% of domestic applicants are admitted. In the *Department of Civil Engineering*, 43% of international and 68% of domestic applicants are admitted (Figures 4.3, 4.4). This indicates that although the number of international applicants is much higher in these programs, the decision-making faculty try to have a balance in the number of international and domestic students who eventually enroll. Posselt (2016) states that "especially in STEM fields, a commonly expressed concern is that international students may be crowding out domestic students" (p. 135). Zhang (2009) found evidence for the displacement of U.S. graduate students in non-STEM fields by the special attention given to international student who have external sources of funding from their own countries.

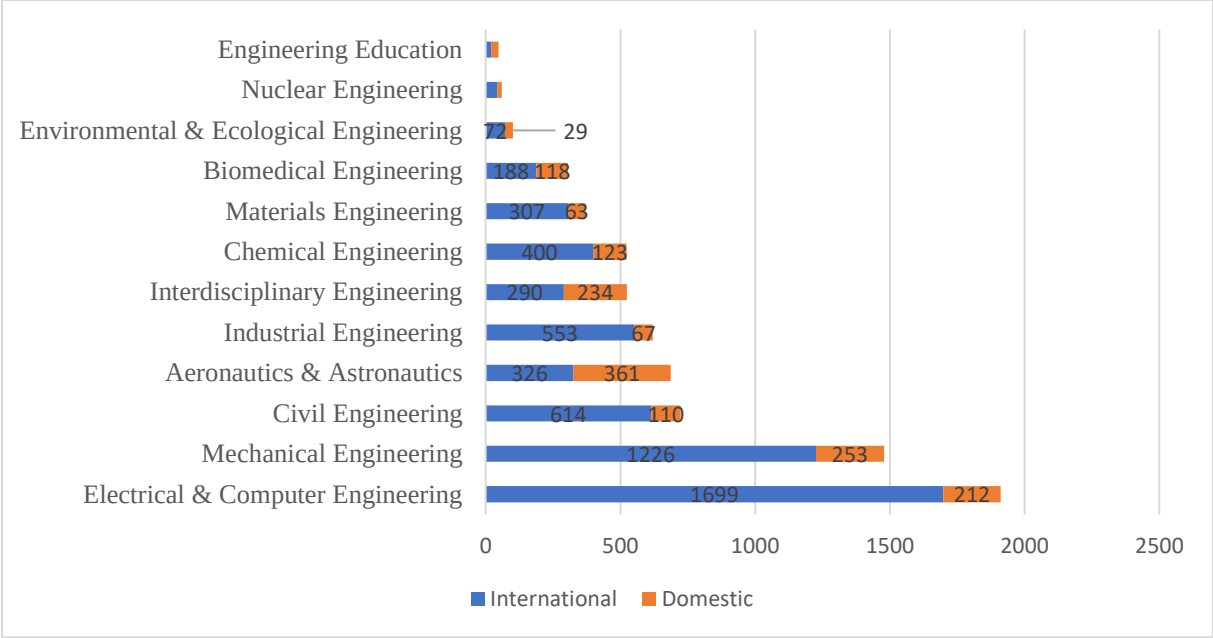


Figure 4.3. Number of All International and Domestic Applicants in College of Engineering Academic Departments – AY 2018/19

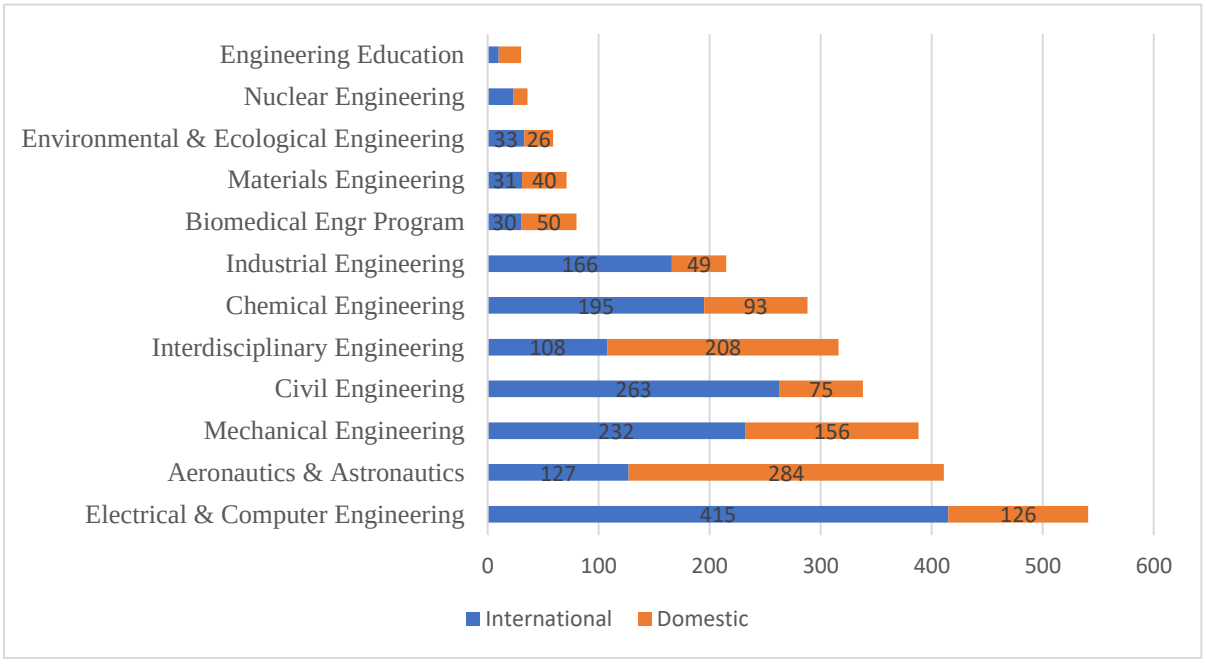


Figure 4.4. Number of Admitted International and Domestic Applicants in College of Engineering Academic Departments – AY 2018/19

The two largest departments in the College of Liberal Arts are *Communication* and *English*, accounting for 48% of all the applicants in this college. In the Communication

Department, only 9% of international applicants are offered admission, which indicates that the programs in this department are selective and competitive for international students. However, in comparison, 75% of all domestic applicants in this department are admitted, indicating that getting into this college is more competitive for international applicants than for domestic applicants. In the English Department, 14% of international and 19% of domestic applicants are admitted (Figures 4.5, 4.6). In the National Foundation for American Policy report, there is no information about the number and percentage of international students in non-STEM fields such as programs in English and Communication, but the effort to even out the number of international and domestic admitted applicants that is visible in STEM fields cannot be detected in non-STEM fields which consist of mostly domestic students.

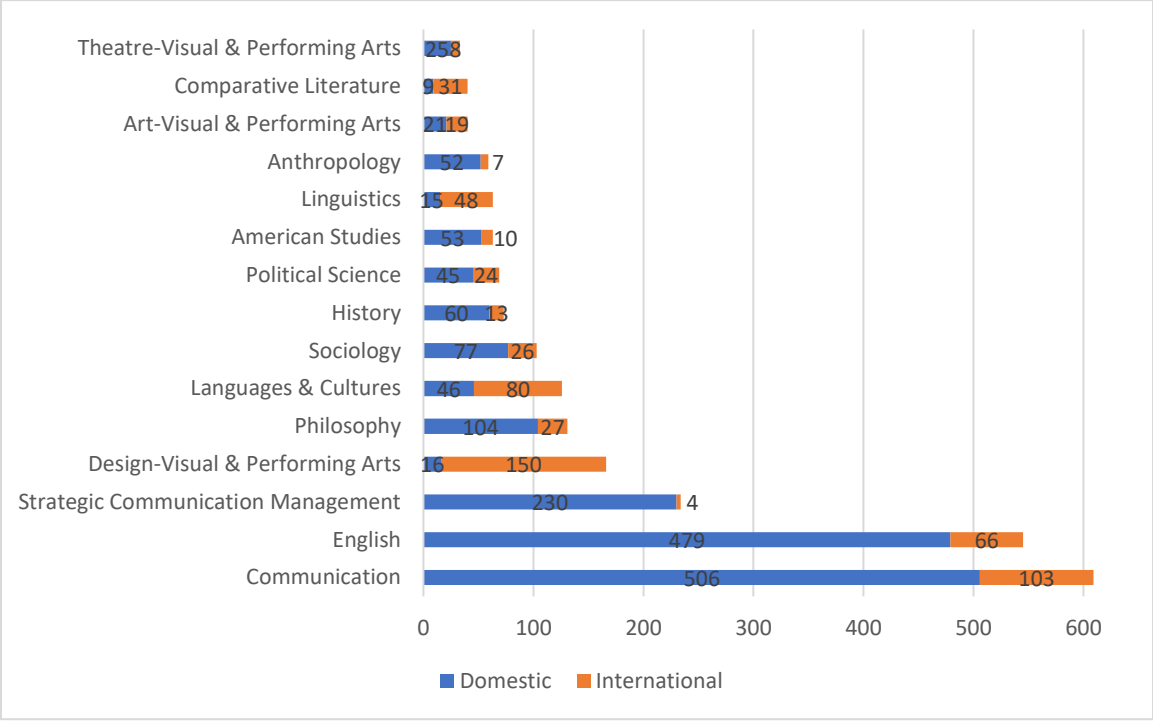


Figure 4.5. Number of All International and Domestic Applicants in College of Liberal Arts Academic Departments – AY 2017/18, 18/19, FALL 2019

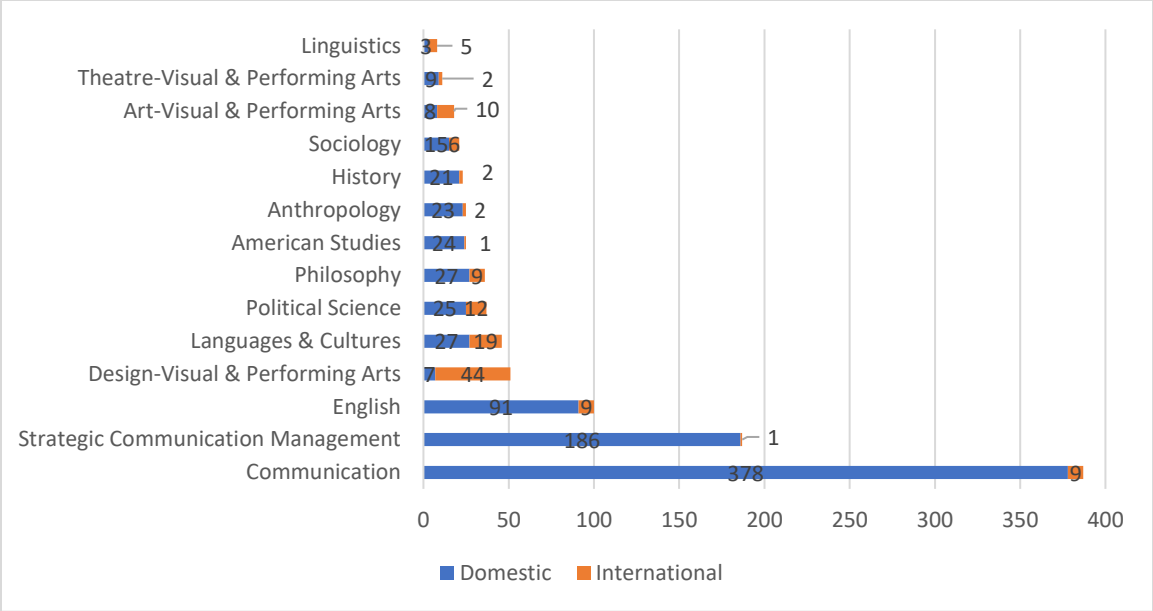


Figure 4.6. Number of Admitted International and Domestic Applicants in College of Liberal Arts Academic Departments – AY 2017/18, 18/19, FALL 2019

The popularity of STEM fields among Purdue's international graduate applicants provides these fields with a wide variety of graduate application profiles to select from. At large public schools such as Purdue, the higher the number of applicants, the more available will be applicants who have both high standardized test scores and the specialized field knowledge necessary to be successful in their graduate programs. Therefore, when making recommendation about student selection based on language proficiency test scores, the fact that not all departments and programs have large numbers of international applicants from which to make ideal selections must be taken into consideration. The nature of a program and the extent to which it is language-heavy might both affect the number of international applicants and limit decision-making faculty's ability to find qualified international students who have a good chance of successfully carrying out the academic tasks required at the graduate level. This could be why in the College of Liberal Arts, only 9% of all admitted applicants are international.

4.2 Language Proficiency Profiles

Graduate programs in the U.S. have been increasingly attracting international students. International graduate students benefit the United States economy by bringing their professional skillset and specialized knowledge with them and contributing to science and academia by

becoming future professors, scientists, and researchers (Posselt, 2016). According to NEAP (2017):

> Without international students, the number of students pursuing graduate degrees (master's and Ph.D.) in fields such as computer science and electrical engineering would be small given the size of the U.S. economy. In 2015, at U.S. universities there were only 7,783 full-time U.S. graduate students in electrical engineering, compared to 32,736 full-time international students. Similarly, in computer science, in 2015, there were only 12,539 full-time U.S. graduate students compared to 45,790 international graduate students at U.S. universities (p. 1).

According to NEAP (2017), approximately 50% of all international students are either from China or India (Table 4.2).

Table 4.2.Top Ten Countries of Origin for International Students (2015/16) – Adapted from Institute of International Education. (2016). Open Doors Report on International Educational Exchange. Retrieved from http://www.iie.org/opendoors

Rank	Place of Origin	Number of International Students – 2015/16	Top Fields of Study
	WORLD TOTAL	1,043,839	Bus./Management, Engineering, Math/Computer Science
1	China	328,547	Engineering, Math/ Computer Science, Bus./Management
2	India	165,918	Engineering, Intensive English, Bus./Management
3	Saudi Arabia	61,287	Bus./Management, Fine/Applied Arts, Social Sciences
4	South Korea	61,007	Bus./Management, Intensive English, Other
5	Canada	26,973	Bus./Management, Engineering, Health Professions, Social Sciences
6	Vietnam	21,403	Bus./Management, Intensive English, Other
7	Taiwan	21,127	Bus./Management, Engineering, Fine/Applied Arts
8	Brazil	19,370	Engineering, Bus./Management, Other
9	Japan	19,060	Bus./Management, Intensive English, Other
10	Mexico	16,733	Bus./Management, Engineering, Other

At Purdue University, the international student body comprises 20.4% of the total number of enrolled students. International undergraduate students comprise 13.8% (4651) of the total undergraduate body. A total of 4,434 international graduate and professional students represent 40.7% of all students at this level of study. International graduate students from China and India comprise a cumulative percentage of 56.5% of all international graduate students. Countries immediately followed by China and Indian are South Korea, Taiwan, Colombia, and Malaysia.

Table 4.3. Top Ten Countries that Represent the Student Body at Purdue University in 2018 . Retrieved from https://www.iss.purdue.edu/Resources/Docs/Reports/ISS_StatisticalReportFall18.pdf

Country	Students	% of Total	Change from 2018
China	3103	32.2%	-6.3%
India	2025	22.3%	2.3%
South Korea	685	7.5%	6.7%
Taiwan	350	3.9%	15.9%
Colombia	166	1.8%	1.2%
Malaysia	158	1.7%	-12.7%
Brazil	143	1.6%	28.8%
Saudi Arabia	114	1.3%	21.3%
Turkey	114	1.3%	4.6%
Indonesia	107	1.2%	-7.8%
Other	2120	23.3%	10.1%
Total	9085	100%	1.7%

The language backgrounds of Purdue's admitted applicants in each college were analyzed. While Chinese students are present in all Purdue graduate colleges, the most diverse college in terms of language background is the College of Science with only 37% Chinese, 8% Indian, and 55% other languages (Figure 4.7). College of Engineering is the college with highest number of Chinese and Indian admitted applicants (a cumulative 77%, see Figure 4.8).

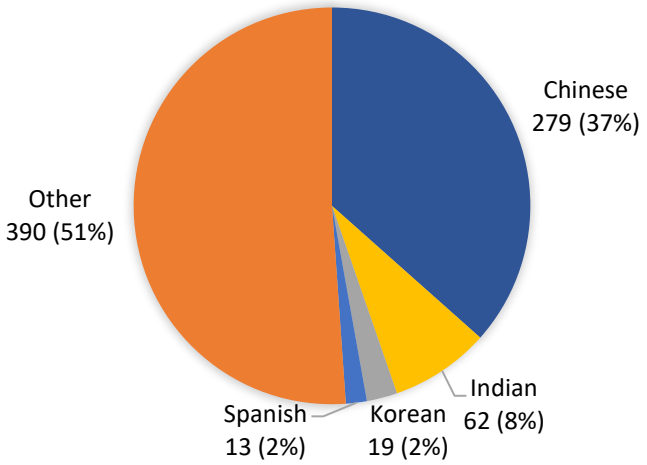


Figure 4.7. Language Backgrounds of Admitted Applicants – College of Science 2018/19

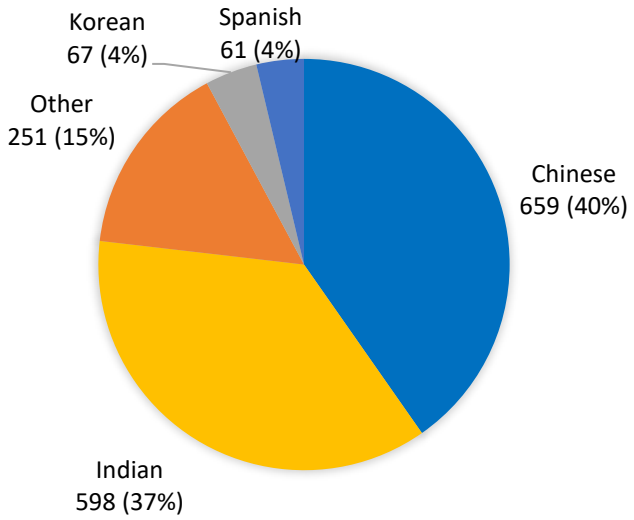


Figure 4.8. Language Backgrounds of Admitted Applicants – College of Engineering 2018/19

Posselt (2016) discusses the ambiguities involved in reviewing international student applications. The fact that graduate applicants come from various language and cultural backgrounds further complicates the process of holistic file review. One of the uncertainties graduate school officials and the deciding faculty have is whether applications from different linguistic and educational backgrounds should be reviewed differently. "How well do indicators of English skills in an application correspond to practical fluency?" asks Posselt (2016, p. 137). This is a very important question, the answer of which could be different depending on the applicant's L1 background.

When discussing the ambiguities encountered when reviewing Chinese and Indian international applicants' graduate application files, it is important to take into account the difference between the status of English in China and India. English as the medium of instruction (EMI) is more prevalent in India than China. EMI refers to the use of English to teach academic subjects in countries where the first language of the majority of population is not English. While English is often a foreign language (FL) in China, the frequent use of English in Indian academic institutions that implement EMI makes the status of English and the experience of learning English much more different for Indian students. After independence from British rule in 1947, due to the existence of many regional languages, the Indian government gradually introduced English as a medium of instruction for maintaining international relations (Sanyal, 2019). In India, the medium of instruction varies among English, Hindi and other official languages, but almost all private schools prefer English, and government (primary/secondary education) schools

tend to go with either English or Hindi. The medium of instruction in colleges and universities of India is always either English, Hindi or a regional language. On the other hand, English is usually taught as a foreign language in China rather than being used as the medium of instruction. Despite the fact that English is introduced to Chinese students starting at Grade 3, the instruction often consists of the rote learning of vocabulary and grammar rules rather than focusing on developing learners' communicative skills (McPherron, 2017).

Test-taking culture can have an effect on students' performance on standardized language proficiency tests. The National College English Test (CET) is a large-scale standardized exam administered by the Ministry of Education in China. According to Gu (2018): "The fundamental purpose of the CET is to comprehensively evaluate English education in Chinese colleges and universities. The test assesses students' English proficiency against the teaching goals prescribed by the Ministry of Education" (para.3). The fact that most institutions in China require the passing of CET for students to qualify for a degree, English language instruction in China adopts a teach-to-test approach which include instruction on test-taking strategies and test-wiseness.

Test-wiseness can be considered a test-taking strategy that could lead to error in the measurement of receptive language skills. Bachman and Cohen (2002) discuss three major sources of variability in language tests, namely, individual differences, test-taking strategies, and the effect of test-tasks on performance on language tests. As early as 1980s, research into test taking strategies, especially those used in multiple-choice reading comprehension items, became popular. Nevo (1989) states that test developers' assumptions regarding what construct their test is measuring might be different from what the test is actually measuring due to the use of test-taking strategies by test-takers; that is, what may be measured is the extent of a test taker's test preparation, rather than language proficiency. Nevo (1989) found evidence for the transfer of contributing test-taking strategies from the first language to the target language, in line with some previously published studies (e.g. Alderson, 1984; Cohen, 1984; Dollerup, Glahn, & Rosenberg-Hansen, 1982). The reading section of a language test is generally more problematic and susceptible to test-wiseness due to the nature of its item types (usually multiple-choice). One of the important research studies related to reading test items and strategy use in reading tests in the 80's and 90's is Anderson, Bachman, Perkins, & Cohen (1991) which investigated the combined interaction between test-taking strategies, item content, and item performance using the think aloud protocol, test content evaluation, and the traditional test performance statistics.

The study shed light on how frequently language test-takers use strategies such as guessing, matching the stem with a previous portion of the text, and referring to time allocations in reading tests. The study revealed that it may be the case that test preparation produces better 'guessers' rather than more proficient language users.

Test-wiseness is a threat to the construct validity of a test (Allan, 1992). In multiple-choice tests, test wiseness can be defined as the ability to identify and use "the cues related to absurd options, similar options, and opposite options" (Cohen 2006, p. 320). Research into the influence of test-wiseness strategies on test performance is not easy since test-wiseness is a tacit trait which is hard to observe and measure. Haiyan & Rilong (2016) investigated the effect of test-takers' use of test-wiseness strategies in Chinese EFL learners' reading test performance. They measured test-wiseness using a questionnaire right before an English achievement test and reported that successful test-takers used test-wiseness strategies no more significantly than unsuccessful test-takers. Their results suggested that "the bias against Asian EFL learners, especially Chinese EFL learners, in their test-taking process" is a misconception since test-wiseness did not contribute to these EFL learners' performance on a reading test (p. 68).

ETS, the developer of the TOEFL iBT, argues that they have designed reading comprehension tasks which measure test-takers' academic reading skills without reliance on test-wiseness. In a TOEFL validation study, Cohen & Upton (2006) collected verbal evidence from 32 test-takers from four different L1 groups, Chinese, Japanese, Korean, and Other. The research found that test-takers use an array of test-taking strategies to improve their performance on reading tasks. The six most common strategies used were:

- Go back to the question for clarification: reread the question.
- Go back to the question for clarification: confirms the question or task (except for basic comprehension—vocabulary and pronoun reference items).
- Read the question and then read the passage/portion to look for clues to the answer either before or while considering options (except in the case of reading to learn—prose summary and schematic table items).
- Consider the options and postpone consideration of the option (except for inferencing–insert text4 items).
- Select options through vocabulary, sentence, paragraph, or passage overall meaning.

- Discard options based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure. (Cohen, 2006, pp. 317-318)

Douglas & Hegelheimer (2005) reported similar strategies used by test-takers during TOEFL listening tasks too. Strategies common in answering TOEFL listening items were "working with the response options by reviewing them in order, narrowing the options to the two or three most plausible, and stopping the review of options without considering the rest when one is considered correct", referring "to prior experience with multiple-choice tests or to prior questions or part of a prior question as a guide to selecting a response", and informed guessing when uncertain about the correct answer (Douglas & Hegelheimer, 2005, as cited in Cohen, 2006).

Yang (2000) investigated the role of test wiseness in taking the TOEFL test. The participants of the study were 390 Chinese test-takers who were given the Test of Test-wiseness developed by Rogers and Bateson (1991), followed by the TOEFL test. After the analysis of the Test of Test-wiseness results, 23 of their participants were deemed "test-wise" and 17 "test-naïve". A combination of these participants' verbal reports and their TOEFL test performance were analyzed, and it was found that 48% to 64% of TOEFL Listening and Reading Comprehension items were susceptible to test wiseness. It was also found that test-wise test-takers were more knowledgeable academically and used their discipline-related knowledge to assist them in eliminating options.

Do Purdue graduate students with different L1 backgrounds perform differently on TOEFL subsections? In order to answer this question and see the difference between test score profiles of the two largest L1 groups of graduate applicants at Purdue, i.e. Chinese and Indian, a hierarchical cluster analysis was conducted with graduate applicants in the College of Engineering (Ginther & Yan, 2018). College of Engineering was selected for this analysis because it is the college with the highest number of admitted international students in AY 2018/19, and because it admits almost equal numbers of Indian and Chinese applicants. The hierarchical cluster analysis procedure generally attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that starts with each case in a separate cluster and combines clusters until only one is left. We used agglomeration coefficients to create a scree plot that revealed the number of clusters we have in the language proficiency scores submitted by admitted applicants at the time of the application to the College of Engineering. According to the

scree plot (Figure 4.9), there are three distinguishable clusters in the data. Table 4.4 presents the centroids for each of the three clusters for each TOEFL subskill for Chinese and Indian graduate students admitted to the College of Engineering in academic year 2018-2019.

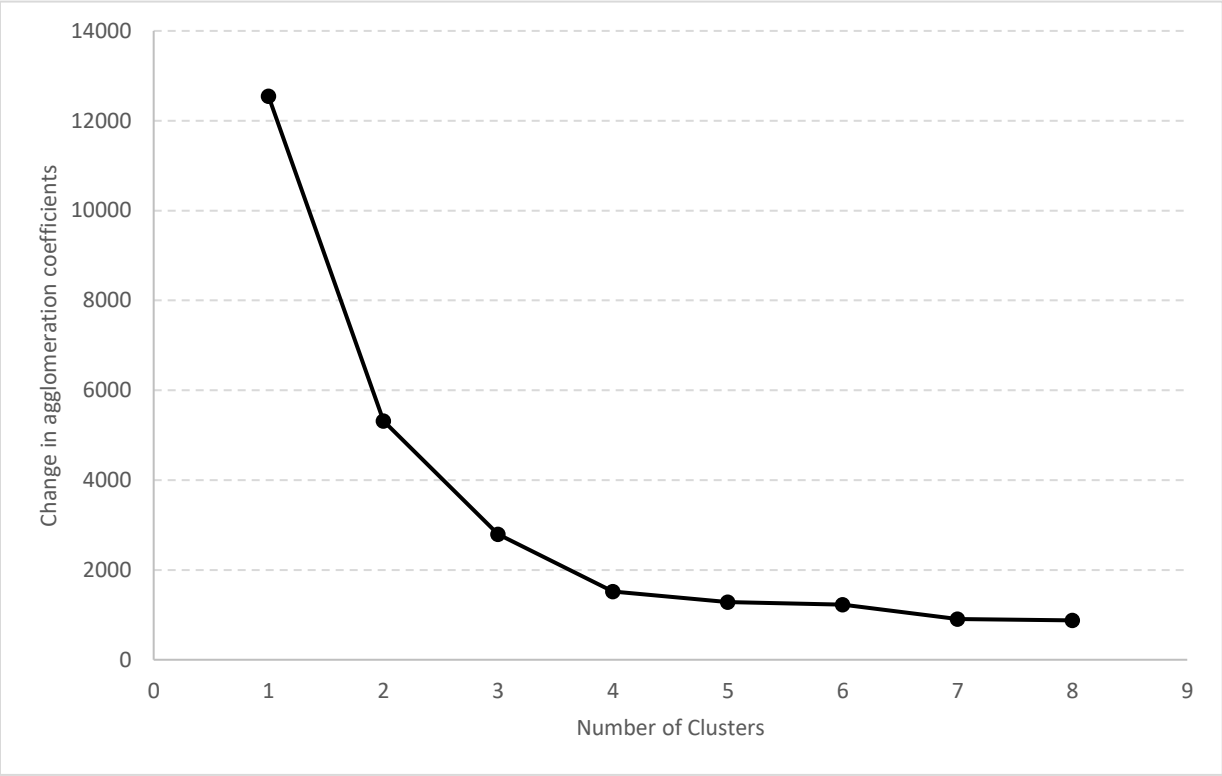


Figure 4.9. Scree Plot for the Change in Agglomeration Coefficients in Purdue's Admitted Applicants' Language Proficiency Profiles – College of Engineering AY 2018/19

As displayed in Table 4.4, in general, the cluster centroids for English proficiency profiles of admitted College of Engineering students indicate the existence of three language profiles: 1) the unbalanced profile (N=565) which belongs to students who have higher scores across the subskills of reading and listening, and comparatively lower scores on speaking and writing. 2) the balanced medium profile (N=231) which represents students who have moderate scores across all four subskills 3) the balanced high profile (N=520) which represents applicants who have high scores across all four subskills. Out of 1316 admitted applicants, only 231 (18%) belong to the 'balanced medium' profile. The majority of applicants are either in the 'unbalanced RL>SW' or in the 'balanced high' profile. The next step is to see if there is an association between language background and membership in one of the three language profiles.

Table 4.4. Cluster Centroids for Subscale Score Profiles

Clusters	Score Profile	N (%)	Reading	Listening	Speaking	Writing
1	Unbalanced (RL>SW)	565 (43%)	28	27	22	25
2	Balanced Medium	231 (18%)	24	22	21	23
3	Balanced High	520 (39%)	29	29	26	28

The cluster analysis results were used to conduct a Chi-square to see if students' profile membership is related to their language background. The purpose of this analysis was to see if students from a specific language background were more likely to have balanced or unbalanced language profiles. The language background groups that I included in the analysis were Chinese and Indian, since the majority (84%) of admitted applicants in Engineering belong to either group. The results of the analysis (Table 4.5) revealed a significant association between language background and belonging to one of the three language profiles. ($\chi 2$=393.03, ρ <.000) with a large effect size (Cramer's V= .39). As it appears in Table 3, 62% of the admitted Chinese applicants have an 'unbalanced RL>SW' language profile, while this percentage for the Indian applicants is only 25%. On the other hand, only 14% of students in the 'balanced high' profile consist of Chinese applicants, whereas 69% of students in this language profile consist of Indian applicants.

Table 4.5. Chi-square Test of Independence between Language Background and TOEFL Profile Membership – Admitted College of Engineering Applicants AY 2018/19

			Profile						
			Unbalanced	Balanced Medium	Balanced High	Total	$\chi 2$	Sig.	Cramer's V
Language	Chinese	Count	348	141	79	568	393.03	.000	.386
		% within Language	61.3%	24.8%	13.9%	100.0%			
		% within Profile	61.6%	61.0%	15.2%	43.2%			
	Indian	Count	140	29	370	539			
		% within Language	26.0%	5.4%	68.6%	100.0%			
		% within Profile	24.8%	12.6%	71.2%	41.0%			

ρ<.05

The findings of the previous analyses are also evident in the graphs below (Figures 4.10, 4.11) which compare the speaking and reading score distributions of Chinese and Indian

applicants. While for speaking, the majority of Indian applicants have TOEFL scores above 24, the majority of Chinese applicants fall in the '21-23' category with very few in the two categories at the higher end of the distribution. However, both Chinese and Indian applicants have a similar trend for reading, scoring very high on this subskill. This adds evidence to the findings that most of the Chinese applicants belong to the 'unbalanced' language profiles with reading, listening scores higher than speaking, whereas most Indian applicants have 'balanced high' profiles with high scores across all four skills.

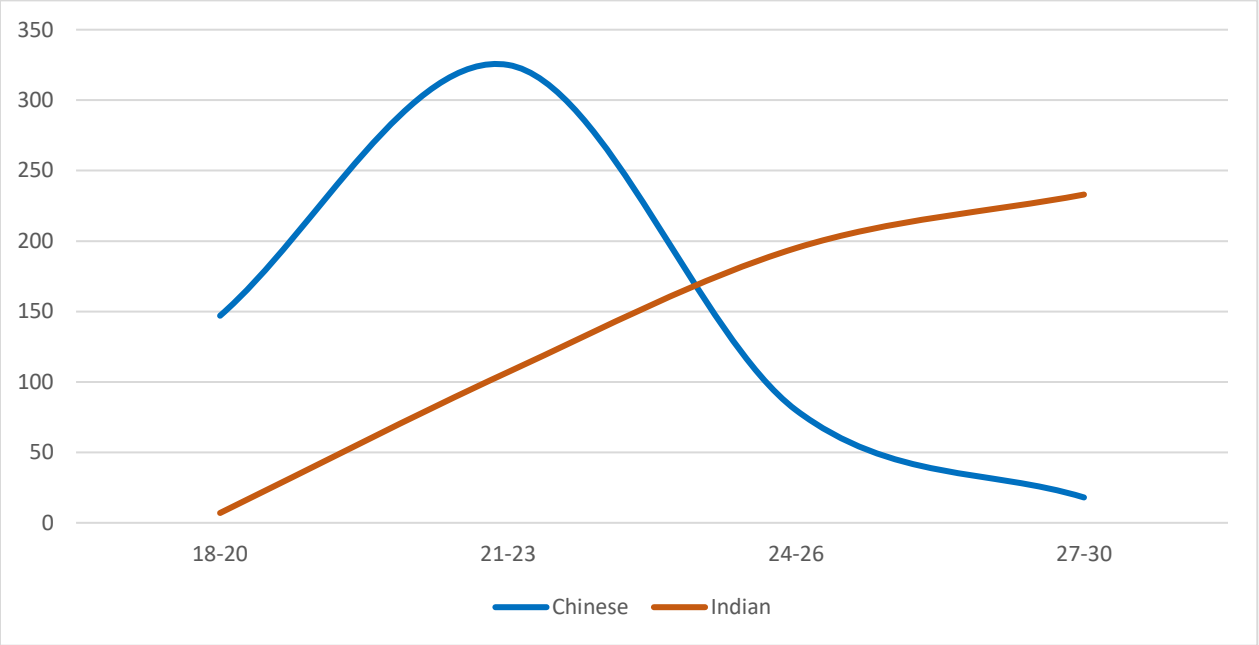


Figure 4.10. TOEFL Speaking Score Distribution across Language Background – College of Engineering AY 2018/19

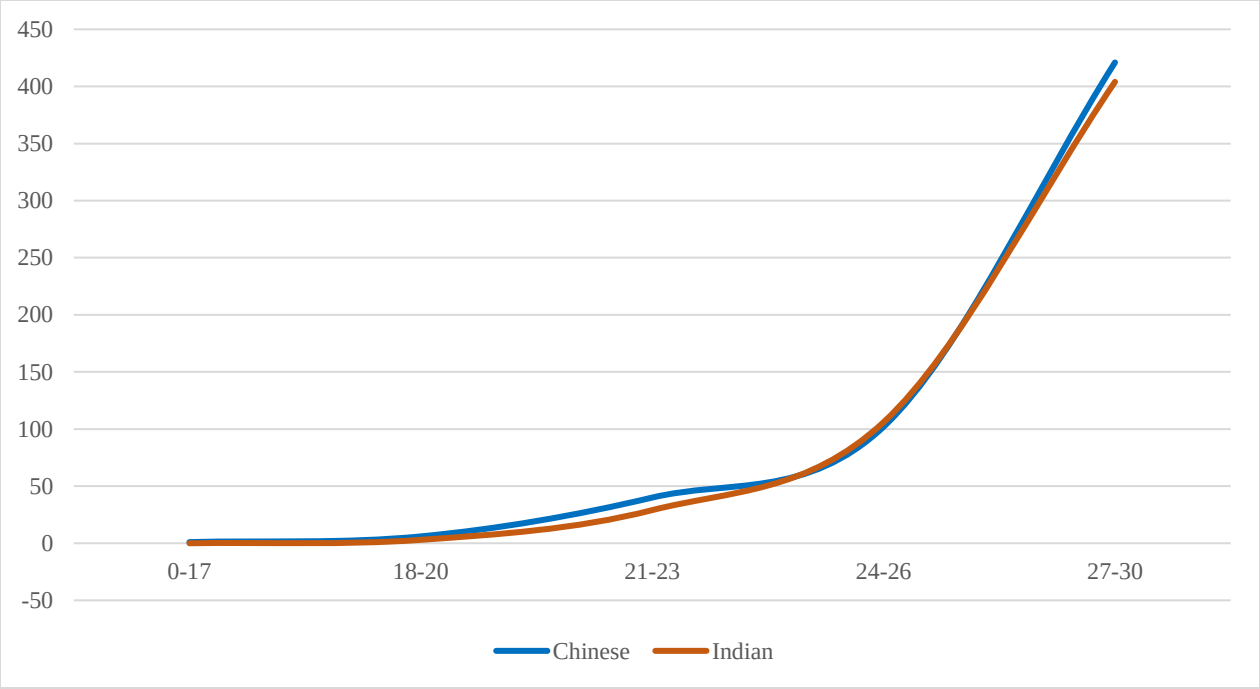


Figure 4.11. TOEFL Reading Score Distribution across Language Background – College of Engineering AY 2018/19

Purdue's Indian and Chinese graduate applicants have similar score patterns in the reading section of the TOEFL, whereas they display a quite different performance on the speaking section. Bridgeman, Cho, & DiPietro (2016) studied uncovered variations in linguistic subgroup differences and different language proficiency profiles among the four skill areas. They conclude that "when the large contingent of Chinese students is separated from the other students, the reading score appears to be important for non-Chinese students but virtually worthless for students from China" (p. 316). They found that the scores from reading and listening sections of the TOEFL gain value only when the data for Chinese students is removed and separated from the rest of the student admissions data. They argue:

> The message for admissions officers then changes from "ignore reading and listening scores for Chinese students" to "pay especially close attention to Chinese students with a large discrepancy between their receptive and productive test scores," especially because these students with large discrepancies do not seem to do very well academically (p. 316).

The findings of the current study are in line with the findings of Bridgeman et al. (2016). The unbalanced language proficiency profile of Chinese applicants clearly illustrates the importance of separately analyzing language proficiency data for different linguistic subgroups.

Bridgeman et al. (2016) found little evidence for the speaking subskill predicting the success of engineering students, but when they split the sample into subgroups of Chinese and Indian, the speaking skill became a very important predictor of success for both groups with an adjusted correlation of .55 with students' grades.

The test-taking culture of China can be one factor contributing to their unbalanced profiles. The competitive exam system in China can be traced back to the imperial exams during the Han Dynasty (AD581–618). Teaching to test has become popular with all the competitive exams which serve as placement tools in various stages of Chinese students' lives (Xiao, 2017). Test-taking strategies, especially the ones that include option elimination, can be taught and learned, which makes the multiple-choice items in the reading and listening sections of the TOEFL susceptible to test-wiseness which is considered a construct irrelevant practice effect (Yang, 2000). Yeom & Jun (2020) compared four different test-wiseness strategies across language proficiency levels and found evidence for the extensive use of 'option elimination' strategy by intermediate proficiency level test-takers.

The extensive use of test-taking strategies in a test-obsessed culture can be the reason for the appearance of such unbalanced language proficiency profiles among Chinese graduate applicants. Cohen (2009) divides test-taking strategies into two categories: test management strategies and test-wiseness strategies. Test management strategies involve strategies such as reading the instructions carefully or being mindful of the time during timed tests, whereas test-wiseness strategies involve guessing the right answer using clues that signal the correct answer. For instance, choosing the longer choice in a multiple-choice question, or eliminating items based on the general world knowledge or because they are not grammatically correct when inserted in the prompt are considered test-wiseness strategies. Cohen (2009) believes that although the use of test-wiseness strategies is generally frowned upon, "survival is the name of the game" and students will attempt anything that would "help them get through a test as effectively as possible" (1:48).

While test management strategies are part of the test-taking process assumed by the test-giver, test-wiseness strategies can be a source of measurement error. Research suggests that test creators, such as the ETS, must be concerned about the construct validity of their tests since test-wiseness strategies can be considered as one source of error in measurement in TOEFL reading and listening sections (Yang, 2000). Although Cohen (2009) states that we still lack a theory

accounting for test-taking strategies, the extensive research on this topic has led to a consensus about what test-taking strategies actually are. Students' use of test-wiseness to improve their performance without possessing the necessary linguistic skills can be a source of invalidity. However, there is still a controversy regarding the effect of test-taking strategies on the construct validity of a test. Some studies show that only highly proficient students can successfully use test-taking strategies to improve their performance, and some other studies show the opposite (Al Fraidan, 2011; Cohen, 2006).

No matter the reason, the frequent appearance of unbalanced language proficiency profiles with the receptive skills much higher than the productive skill can be problematic in the process of graduate student selection. The results of the analyses indicate the importance of carefully considering subskill proficiency scores in addition to the total score when admitting students because a falsely inflated reading or listening score will also inflate the total score and lead to overestimation of the graduate applicant's language proficiency. Chinese students who come in with lower productive skills will need further support from the university, especially if they need to be placed into teaching assistantship positions.

Reliance solely on the TOEFL total score rather than subskill scores is evident in many pre-admission and post-admission processes at Purdue and other higher education institutions. While there are many institutions which select students based on only one cut score for TOEFL total, setting low admission standards for subskill scores, as in Purdue, will also lead to the appearance of unbalanced language proficiency profiles in the pool of applicants being considered for admission. An example of a post-entry language proficiency screening practice which might get affected by an inflated TOEFL total score is Purdue's Department of Food Science writing proficiency screening. The Food Science Graduate Handbook 2019-2020 states:

> If a student whose native language is English does not satisfy the … TOEFL requirements, the Graduate Committee can administer a screening test in written English to determine if additional training is needed to become proficient in English composition. If the TOEFL score is 80 (iBT based exam), the student will be required to take the written English screening test. (p. 14)

Considering the TOEFL total score while ignoring scores on the writing section can be problematic, especially among the Chinese graduate student population. Students who have an inflated total score due to high scores on reading and listening will never be screened for writing proficiency in the Department of Food Science.

4.2.1 Implications for Language Assessment Literacy (LAL)

Language assessment literacy is generally viewed in the literature as a set of skills and the knowledge to use the existing assessment methods, develop suitable assessment tools to assess a construct of interest, and analyze the data generated from a test (Inbar-Lourie, 2008; Pill & Harding, 2013; Stiggins, 1999). However, in the context of admissions, the definition of LAL is different to the extent that it cannot be called *Assessment Literacy* anymore. Through the "application of conceptualizations of *literacy* from other fields to LAL", Pill & Harding (2013) denied the binary classification notion of *literacy* and *illiteracy.* Instead, they argue that *literacy* can be classified into *nominal literacy, functional literacy, procedural and conceptual literacy* and *multidimensional literacy* (p. 383):

Table 4.6. Five Stages of Literacy in LAL. Adapted from Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. Language Testing, 30(3), 381–402.

Illiteracy	Ignorance of language assessment concepts and methods
Nominal literacy	Understanding that a specific term relates to assessment, but may indicate a misconception
Functional literacy	Sound understanding of basic terms and concepts
Procedural and conceptual literacy	Understanding central concepts of the field, and using knowledge in practice
Multidimensional literacy	Knowledge extending beyond ordinary concepts including philosophical, historical and social dimensions of assessment

The definition for each stage of literacy is transformed by Pill & Harding (2013) to fit into the field of LAL. Based on these definitions, progress into which stage of assessment literacy is necessary for the decision-making faculty in graduate admissions to make informed decisions about test-takers based on their language proficiency scores? Since the term *Assessment* Literacy focuses on measurement, which can be excessive in the context of graduate admission, I redefined the stages of literacy in the context of admissions for *Language Proficiency Literacy (LPL)* rather than *Language Assessment Literacy* as follows:

Table 4.7. Five Stages of Literacy in LPL transformed from Pill & Harding (2013)

Illiteracy	Ignorance of the characteristics of language proficiency test-takers and meaning of language proficiency test scores and cut scores
Nominal literacy	Basic understanding of the characteristics of language test-takers and what language proficiency test scores and cut scores represent in the context of admissions, but may indicate a misconception
Functional literacy	Sound understanding of the characteristics of test-takers and what level of proficiency language proficiency scores and cut scores represent and using that knowledge in decision-making
Procedural and conceptual literacy	Understanding central concepts of the field of language assessment, and using knowledge in decision-making
Multidimensional literacy	Knowledge extending beyond ordinary concepts including philosophical, historical and social dimensions of language assessment

Based on the findings of Ginther & Elder (2014), we can conclude that the majority of admissions decision-makers fall in either of the first two stages: Illiteracy or Nominal Literacy. Most of the decision-making faculty either ignore the meaning of language proficiency test scores and cut scores set by the Graduate School or they have minimal knowledge including misconception. If the involvement of the language testing community in LPL development for graduate admissions is sought, its effort must be towards providing decision-makers with the knowledge/skills necessary to reach the third stage of LPL: Functional Literacy. While the fourth and fifth stages of LPL are not necessary for admissions decision-makers, familiarization with the linguistic characteristics of graduate applicants, meaning of language subskill scores, nature of the constructs being measured, and meaning of cut scores set by Graduate Schools and departments are important for reaching a functional level of language proficiency literacy. Any misconceptions about language proficiency test scores and cut scores must be eliminated and replaced by functional information about the use of language proficiency scores in graduate admissions.

To help faculty gain functional language proficiency literacy, it is also important to eliminate any misconceptions they might have about the minimum cut scores required by the graduate schools. At Purdue, the minimum language proficiency scores set by the Graduate School for being considered for admission into various academic programs are quite low (Writing 18, Speaking 18, Listening 14, Reading 19, Total 80 for the TOEFL and Reading 6.5, Listening 6.0, Speaking 6.0, Writing 5.5 for the IELTS). Ginther and Elder's (2014) admissions decision-making respondents differed in their perception of the cutoff scores set by the graduate school for admission to graduate programs; at Purdue, "52% of the respondents … indicated that

they understood the language proficiency requirements set by the university as minimal English-language proficiency requirements, while 38% of the respondents indicated that they believed the requirements represent adequate English-language proficiency" (p. 14). There were even 3% who believed the cutoff scores represented an "advanced" level of language proficiency. At Purdue, increasing graduate committee members' awareness of the fact that Graduate School's language proficiency cut scores represent only minimally adequate language proficiency is an important step towards increasing decision-makers' functional language proficiency literacy. One reason for setting low cutoffs for language proficiency could be due to Graduate School's attempt to adhere to the Holistic File Review and to give the decision-makers the opportunity to consider almost all applicants for admission into their programs. However, due to the different conceptions/misconceptions decisions-makers might have about language test scores and the cutoffs, it is important to convey the information necessary for them to gain functional LPL and to eliminate any misbeliefs.

One of the important pieces of information that needs to be communicated with the decision-making faculty is the linguistic characteristics of various test-taker groups and the effect of those characteristics on the extent of student support needed for improvement. The findings of this study suggest that students from various language backgrounds can display language proficiency profiles that are quite different from one another. As the largest group of applicants to Purdue Graduate School, the possibility of Chinese applicants' belonging to an unbalanced language proficiency profile requires attention to their subskill scores, especially when the admitted applicant is expected to perform roles that require higher English communicative skills, such as teaching a class or the lab section of a course. While a composite total test score is important in the admission of Chinese graduate applicants, high receptive skill scores can contribute to a high overall score, which could be misleading. For instance, A student with a balanced language proficiency profile with the score of 25 in each of the four subskills on the TOEFL has the same total score as a student with an unbalanced profile with the scores of 28 in reading and listening, a score of 23 in writing and a score of 21 in speaking. These two students will perform quite differently in an academic communicative context. This is also evident in Figure 4.12 which displays the importance of students' TOEFL speaking scores in addition to their TOEFL total score in passing Purdue's Oral English Proficiency Test (OEPT) which is a post-entry speaking test administered to prospective graduate teaching assistants. While 98% of

OEPT test-takers who have TOEFL total scores above 99 and speaking scores above 24 pass the OEPT, only 56% of those with TOEFL total scores above 99 and speaking scores between 22 and 25 pass the OEPT. Only 32% of those with TOEFL total scores above 99 and speaking scores below 21 pass the OEPT. While I do not suggest that only students with high speaking subskill scores should be admitted, I believe it is important on the part of the decision-making faculty to give meticulous thought to the roles each student is expected to play in the specific academic setting of their graduate program and the language skills necessary to perform those roles.

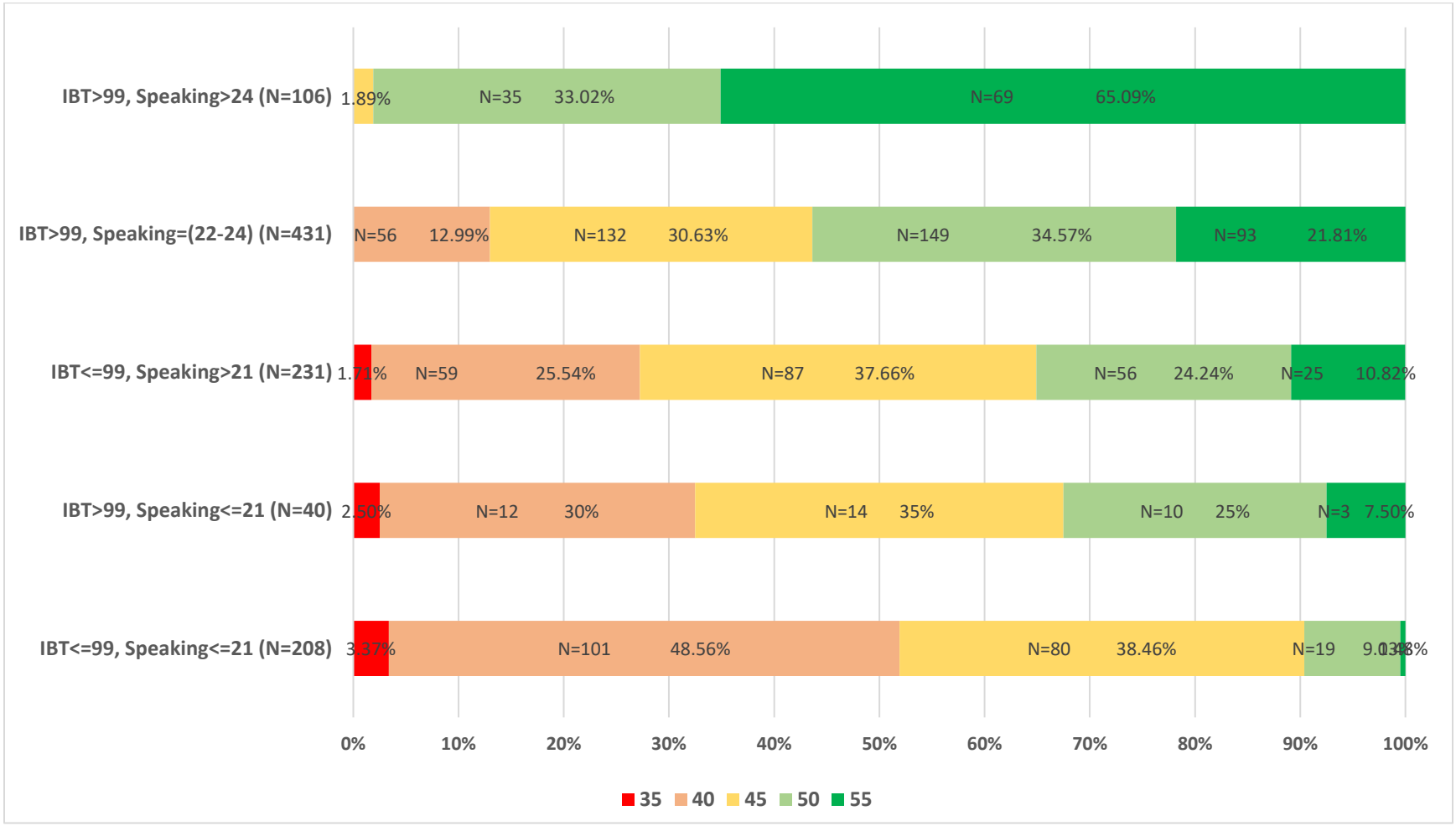


Figure 4.12. The Percentage of Students in Five different TOEFL Speaking Score Categories in Each OEPT Score Level

4.3 TOEFL Score Distributions

One of the important pieces of information which can enable language testers to gain admissions literacy before they can offer assessment literacy in the context of admissions is the distribution of applicants' language proficiency test scores across admission status. The difference between distribution trends across admission status can be informative in detecting the weight language proficiency test scores carry in the process of student selection, and what the ideal distribution can look like. Figure 4.13, Figure 4.14, Figure 4.15, and Figure 4.16 display the distributions of the four TOEFL subskill and total scores of international applicants by admission status for the academic year of 2018/19 at Purdue's Engineering Department. We can see that the score distributions for reading, listening, and writing are more skewed to the right and have a steady increase with the increase in scores as compared to the distribution of speaking for all the three groups (i.e. admitted, rejected, and matriculated), indicating that students generally score higher in reading and listening than speaking. The trends for the three groups do not look drastically different in any of the graphs. Admitted and rejected applicants' TOEFL speaking score distribution curves are closer to a normal distribution curve than any other subskill with '21-23' being the most populated score range. However, the curve for matriculated students almost flattens out after the '21-23' category, indicating that students who enroll at Purdue College of Engineering have a wide range of speaking scores from intermediate to advanced. For reading and listening, admitted applicants' scores are heavily skewed to the right, indicating that most of the students who are offered admissions have very high scores on these two subskills while having lower scores on speaking. It is important to notice that there are no clear differences between the distributions of admitted, rejected, and matriculated students.

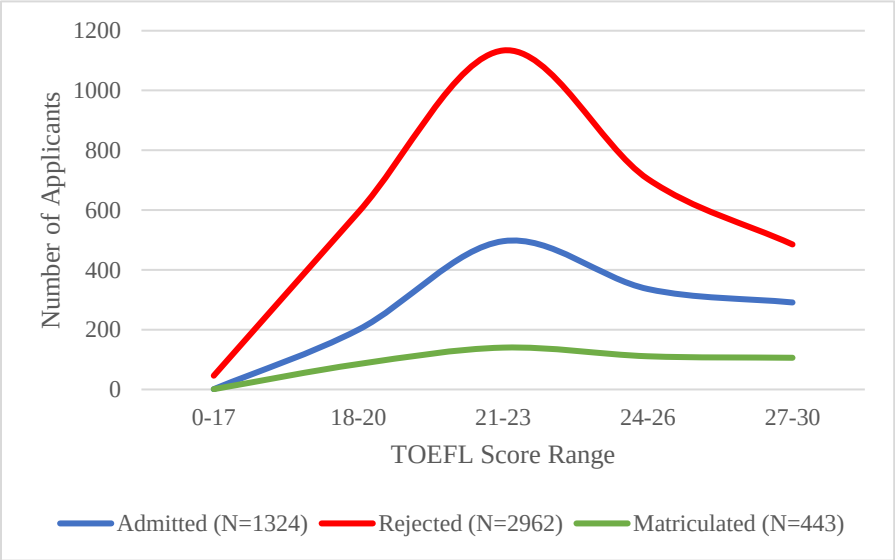


Figure 4.13. TOEFL Speaking Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19

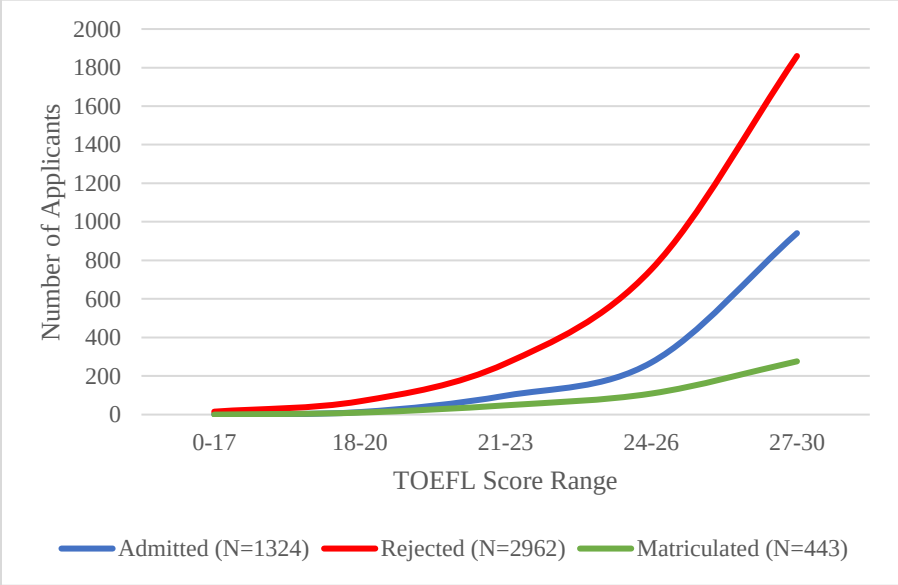


Figure 4.14. TOEFL Reading Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19

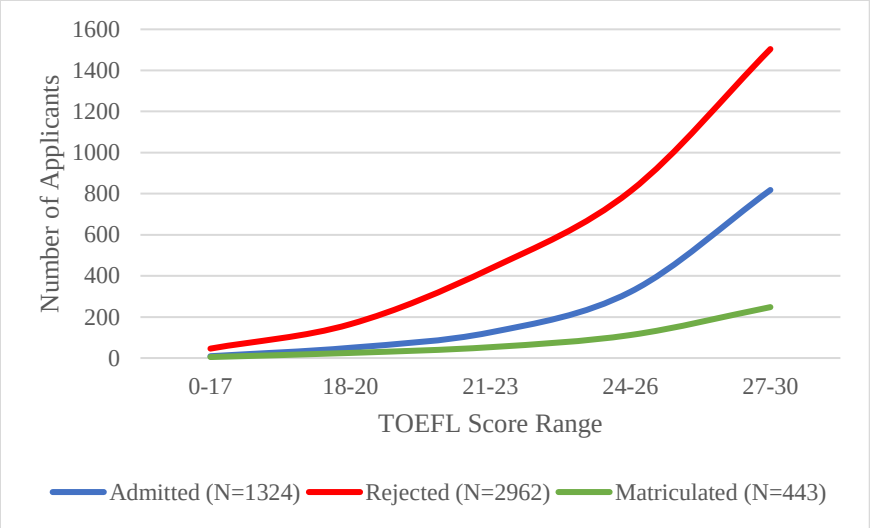


Figure 4.15. TOEFL Listening Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19

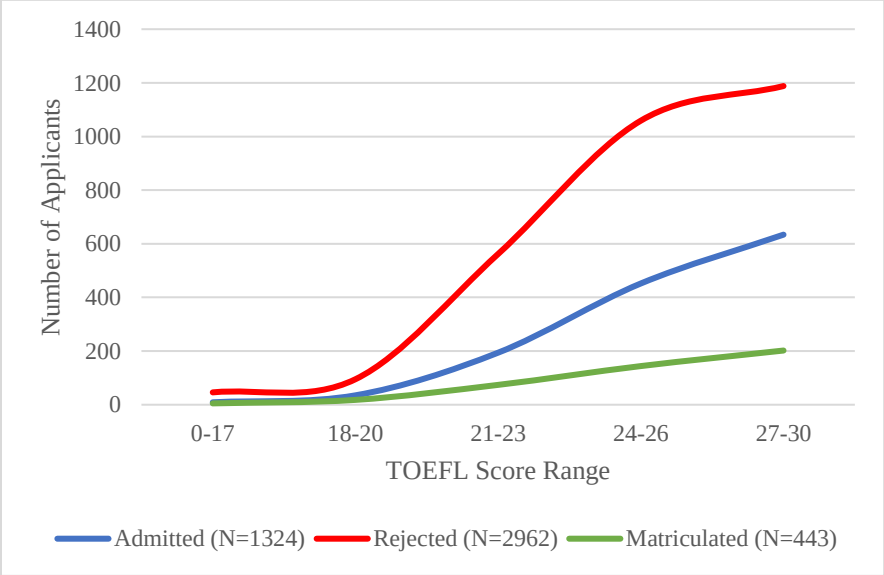


Figure 4.16. TOEFL Writing Score Distribution of Admitted, Rejected, and Matriculated Applicants in College of Engineering – AY 2018/19

4.3.1 Implications for Language Assessment Literacy

There are two pieces of information in the graphs above that are considered valuable to test users. One is the general trend of the distribution for each subskill score and the other is the difference between the distribution patterns of admitted, rejected matriculated students. The drastic difference between the distribution pattern of the speaking subskill and the other four

subskills indicates that there are many students who are admitted with higher scores in reading, listening and writing than speaking which can be problematic if students are placed in roles that require them to be professionally involved in oral communication with others, especially with undergraduate students. Ginther (2003) discusses a history of what was called the *foreign TA problem* in the 1980s when "the… undergraduate difficulties with international teaching assistants led to the establishment of mandates, passed by state legislatures…requiring that the oral English language proficiency of prospective ITAs be certified before those students would be allowed to have direct contact with undergraduates" (p. 59). These mandates required the use of an oral English proficiency test to assess communicative proficiency and further resources to ensure at least a minimally adequate oral English proficiency before ITAs could start performing their teaching roles. The amount of post-entry assessment of oral English proficiency Purdue does each year and the amount of post-entry language support necessary to get prospective ITAs ready to teach are greatly affected by the amount of attention paid to subskill scores when selecting graduate students in the first place.

The lack of difference between the three groups (i.e., admitted, rejected, matriculated) in the subskill TOEFL score distribution graphs can be alarming and could indicate that language proficiency subskill score is not among determining factors when deciding on the admissibility of a graduate student to a graduate program. While it is true that language proficiency is not the only factor determining the decision about a graduate application, in an ideal (yet realistic) setting, we can expect to see a graph that looks like Figure 4.17 for all four subskills and the total score across admitted and reject applicants. In the Engineering department, 40% of rejected applicants have TOEFL speaking scores of 24 and above (N=1189). This finding suggests that admissions committees are not paying attention to or perhaps do not value communicative language proficiency as an important characteristic in a graduate applicant. Specific attention paid to the subskill scores by the decision-makers could result in score distributions that look more closely like Figure 4.17 for all four subskills.

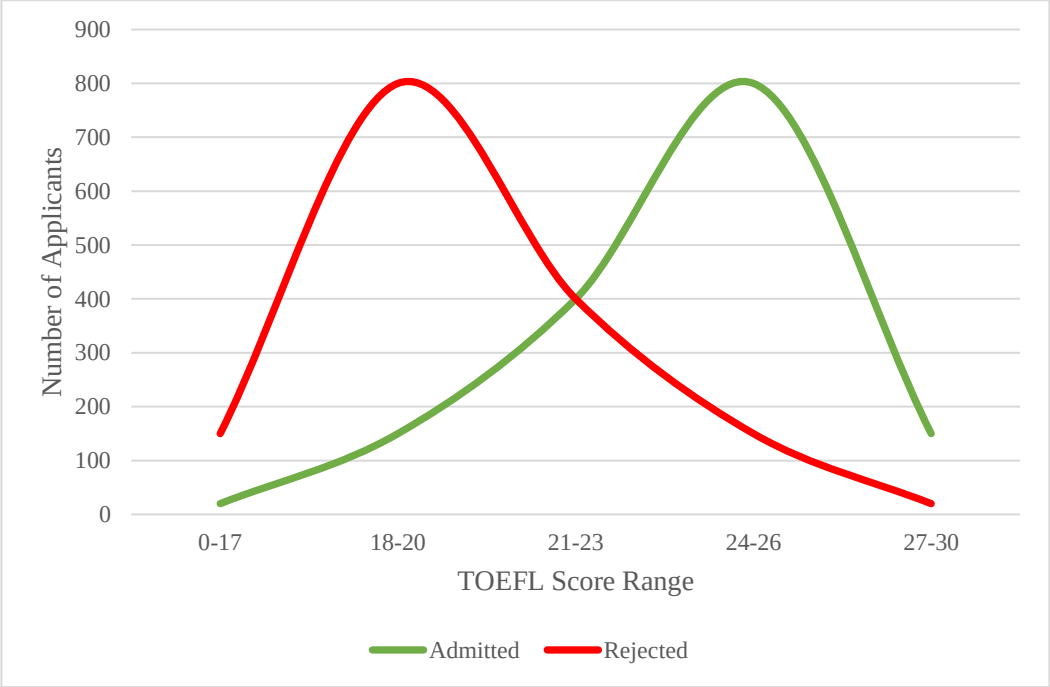


Figure 4.17. A Hypothetical TOEFL Subskill Score Distribution of Admitted and Rejected Applicants in College of Engineering

CHAPTER 5. CONCLUSIONS, LIMITATIONS, AND RECOMMENDATIONS FOR FURTHER RESEARCH

5.1 Summary of the Study Findings and Implications

This study analyzed the international graduate application dataset, which consists of Purdue graduate application information from Fall 2016 to Fall 2020, to investigate the characteristics of graduate applicants with regard to their language proficiency in each college and large department at Purdue University. The study first examined the characteristics of graduate applicants in stem and non-stem colleges and found that in non-STEM Purdue programs, such as programs in the college of Liberal Arts or College of Education, the number of international applicants is lower compared to STEM programs. While the number of admitted and matriculated applicants in STEM programs at Purdue were in line with the general trend in engineering reported by the National Foundation for American Policy, it was hard to find nation-wide enrollment information for non-stem disciplines. The study concluded that the popularity of STEM fields among Purdue's international graduate applicants provided these fields with a wider variety of English proficiency profiles to select from. This has implications for the language testing community when providing LAL development opportunities for admissions decision-makers; when making recommendation about student selection based on language proficiency test scores, the fact that not all departments and programs have large numbers of international applicants from which to make ideal selections must be taken into consideration.

One of the important pieces of information that needs to be communicated with the decision-making faculty is the linguistic characteristics of various test-taker groups and the effect of those characteristics on the extent of student support needed for improvement. Examining the language proficiency profiles of graduate applicants at the College of Engineering, the study found three distinct language proficiency profiles: unbalanced, balanced medium, and balanced high. In the subsequent Chi-square analysis, the study found that the majority of Indian applicants belong to the balanced high profile, indicating that most Indian graduate applicants have high English proficiency test scores across the four subskills, whereas the majority of Chinese applicants belong to the unbalanced English proficiency profile, indicating that they have high scores on receptive skills and lower scores on productive skills. The fact that Purdue's Indian and Chinese graduate applicants have similar score patterns in the reading section of the

TOEFL, whereas they display a quite different performance on the speaking section has implications for the graduate admissions decision-making process; as mentioned in Bridgeman, Cho, & DiPietro (2016): “The message for admissions officers then changes from “ignore reading and listening scores for Chinese students” to “pay especially close attention to Chinese students with a large discrepancy between their receptive and productive test scores” (p. 316).

The study also examined the difference between TOEFL subskill and total score distributions across admission status to look at the weight language proficiency test scores carry in the process of student selection, and what the ideal distribution could look like. The study found that the score distributions for reading, listening, and writing have a steady increase with the increase in scores as compared to the distribution of speaking, which resembles a bell curve for all three groups (i.e. admitted, rejected, and matriculated). This indicates that students generally score higher in reading and listening than speaking. The trends for the admitted, rejected, and matriculated groups do not look drastically different. Two pieces of information that are considered valuable to test users found as a result of these analyses are: 1) The drastic difference between the distribution pattern of the speaking subskill and the other four subskills indicates that graduate matriculated students with unbalanced profiles can encounter problems if they are placed in roles that require them to be professionally involved in oral communication with others, and 2) the lack of difference between the three groups (i.e., admitted, rejected, matriculated) in the subskill TOEFL score distribution indicates that language proficiency subskill score is not among determining factors when faculty members are making admissions decisions.

5.2 Limitations of the Study

One main limitation of the study is its narrow context. This study was conducted in the context of Purdue University, and the findings might not be generalizable to other graduate admissions contexts. The goal of this study was to take a first step in raising admissions decision-makers’ awareness of the use of language proficiency test scores in making informed decisions when selecting graduate students. While the findings of the study are highly informative for the decision-making faculty here at Purdue University, the trends found in this study might not resemble the trends detected in other graduate application data. One of the important assertions of the present study is the uniqueness of graduate application trends across

various contexts and disciplines, and therefore, conclusions about other graduate admissions contexts must be made after similar analyses are conducted in those contexts.

Another limitation of the study was that it was mostly devoted to comparing the linguistic characteristics and trends of Indian and Chinese applicants because they are the largest language groups that apply to graduate programs at Purdue University. Therefore, the comparisons made in this study might not be generalizable to contexts where the majority of graduate applicants come from language backgrounds other than Chinese and Indian. Analyzing the data for more distinct language groups might yield results that are inclusive and informative when it comes to LPL development.

One other limitation of the study was the lack of sufficient time to communicate the findings of the study with the graduate admissions committees of each college/large department to analyze the longitudinal effects of raising admissions decision-makers' LPL. Sending out LPL reports to each college/large department at Purdue about the meaning of language proficiency test-scores and the Graduate School cut scores, and the linguistic characteristics of English test-takers in each college and department is the ultimate goal of this study.

5.3 Recommendations for Further Research

As mentioned before, the current study is a first step in providing the admissions decision-makers with the opportunity to increase their awareness and knowledge about language proficiency test scores and the linguistic characteristics of their graduate applicants. As one of the largest groups of test score users, the decision-making faculty should be familiar with the various language proficiency profiles their graduate students display and know how their students will perform in different roles assigned to them. Therefore, a subsequent goal of the current study is to provide the decision-making faculty with the above-mentioned information. One of the important considerations before planning LPL development opportunities must be the timing of such practices. Faculty members should receive the LPL development information at a time when it is most relevant to their roles as admissions decision-makers. Since the graduate application deadlines differ in various programs at Purdue University (e.g., May 15 for Civil Engineering, January 6 for English, and March 1 for Computer and Information Technology), faculty members in each department must receive the LPL development materials before they

begin the decision-making process, which is after their department's or program's application deadline has ended.

Analyzing the longitudinal effects of providing faculty members with this information can guide our future endeavors to educate this group of test users about the use of language proficiency test-scores in the admissions process. Examining the difference between the language proficiency profiles of admitted graduate students and their TOEFL distribution trends over the course of several years before and after delivering LPL development opportunities will be an invaluable research study.

There are multiple ways by which the above-mentioned information can be delivered to the admissions decision-makers. This includes reports, memos, workshops, trainings, etc. Since faculty members are generally busy, they might be reluctant to spend an extended amount of time to participate in LPL development activities. A study conducted to find the most efficient and effective way of raising faculty members' awareness about language proficiency test scores can be very informative to the language testing community.

As mentioned above, one limitation of the study is that it compared the linguistic characteristics and trends of only Indian and Chinese applicants because these are the largest language groups to apply to graduate programs at Purdue University. While the study ignores the language proficiency profiles of other L1 groups, a study conducted to compare the characteristics of graduate applicants from several distinct language group could present to admissions decision-makers a clearer picture of the effect of linguistic profile differences on academic performance.

APPENDIX A. OEPT HOLISTIC SCALE

Level	*OEPT HOLISTIC SCALE for RATERS the symbol / means* "and or"
55	***MORE THAN ADEQUATE PROFICIENCY for classroom teaching. At least half of items rated 55.*** *Strong skills evident on all items. Little listener effort required to adjust to accent/ prosody/ intonation.* ***Consistently intelligible, comprehensible, coherent, with displays of lexico-syntactic sophistication, fluency and automaticity.*** *Speaker is capable of elaborating a complex or personalized message/argument using a variety of tense/aspect and mood. May show minor fluency or prosodic issues (e.g. occasional misplaced stress, hesitations, filled pauses, occasionally speaks too fast). Any grammar errors are minor (e.g. omission of 3rd pers. sing. present morpheme). Good listening comprehension. Speaker has sufficient range, depth and sophistication of English to communicate successfully in any instructional position.*
50	***ADEQUATE PROFICIENCY for successful classroom communication without support. At least half of items 50 or above.*** *Small amount of listener effort may be required to adjust to accent/prosody/ intonation, but adjustment happens quickly.* ***Consistently intelligible, comprehensible, coherent****. Capable of elaborating beyond the prompt with some detail and specificity. Elaborates coherent messages/arguments. Speaker may exert some noticeable effort, and speed may be variable, but there are some fluent runs and no pattern of disfluencies. Despite minor errors of grammar/vocab usage/stress which do not interfere with listener comprehension, message is coherent and meaning is easy to follow. Some lexico-syntactic sophistication, more than basic vocab usage and syntax, ability to paraphrase.* ***Good listening comprehension.*** *Is currently capable of consistently successful classroom communication without support.*
45	***NOT QUITE ADEQUATE or INCONSISTENT PERFORMANCE ACROSS ITEMS – Majority of items 45. Capable of classroom communication but, due to weaknesses, speaker requires support.*** *Tolerable listener effort required to adjust to accent.* ***Consistently intelligible and coherent.*** *Strengths & weaknesses, inconsistencies across other characteristics of speech or across items. Profiles vary: Responses may require more than a little noticeable effort for speaker to compose, delivery may be slow and hesitant (but not disfluent); Message may be generally clear and expressed fluently, but vocab/syntax may be somewhat basic or often inaccurate; responses/messages may tend to be general/generic rather than specific or detailed; pronunciation/stress/prosody may need refining in order for speaker to be easily understood/followed.* ***Good listening comprehension*** *but may simply repeat information verbatim without paraphrasing****.*** *Has minimally adequate lexico-syntactic resources and fluency necessary for classroom communication and interaction, but requires support to identify weaknesses and improve in order to reach the next level of proficiency required for certification.* **List specific areas that speaker would need to improve in order to be certified**.
40	***LIMITED Language resources/ability to communicate at a level necessary for classroom teaching is limited - Not ready for classroom teaching. Mix of 40 and 45 item scores, or majority 40 with a few 35s, if any. Able to fulfill most tasks****, but weaknesses are obvious. Profiles vary: Consistent listener effort may be needed to follow message. Speaker may be* ***occasionally*** *unintelligible/incomprehensible/incoherent. Grammar and/or vocab resources may be limited. Message may be simplistic/repetitive/unfocussed/ occasionally incorrect. Speaker may have to exert noticeable efforts to build sentences/argument or to articulate sounds. Despite all their shortcomings, these speakers are generally able to get the message across, albeit a simple, incomplete, generic or vague one.*

35	***RESTRICTED Language resources or ability to communicate is RESTRICTED – Likely to need more than one semester of support. <u>Mix of 35 and 40 item scores.</u>*** *Listener may need to exert considerable effort to follow, or may not be able to follow. Profiles vary: Speaker may be more than occasionally unintelligible or incoherent OR may be* ***restricted*** *in several of these areas: fluency, vocabulary, grammar/syntax, listening comprehension, articulation/pronunciation, prosody (includes intonation, rhythm, stress),* ***often resulting in difficult, frustrating or unsuccessful communication. May not be able to adequately fulfill tasks. Not ready for ENGL 620. Explain specific issues that make the speaker unprepared for ENGL 620.***

APPENDIX B. TOEFL IBT SCORE USE AND INTERPRETATION MEMO, 2020

PURDUE UNIVERSITY | Oral English Proficiency Program
COLLEGE OF LIBERAL ARTS

To: Department Heads, Grad Advisors, and OEPP Liaisons

From: April Ginther, Professor of English and Director of the Oral English Proficiency Program

Date: February 6, 2020

Re: TOEFL iBT Score Use and Interpretation

Attached you will find a table showing TOEFL iBT score subscales linked to the Common European Framework of Reference (CEFR) descriptors. The OEPP recommends that graduate admission committees select students who have the higher scores associated with the CEFR B2 level if those applicants are being considered for graduate teaching assistantships (see below and attached).

Applicants who have a TOEFL **total score of at least 100 and 25 on the speaking portion** of the test are likely to pass the OEPT.

Applicants who have at least **100 as a total score and at least 22 on each subscale** are more likely to score a 45 on the OEPT.

Prospective International Teaching Assistants (ITAs) who receive a 45 on the OEPT can teach while being enrolled in English 620. Those who score 40 on the OEPT cannot teach while taking the English 620 course.

The table we have provided also includes TOEFL iBT percentiles for graduate test takers. Please note that the Purdue University Graduate School admissions cut scores are at the lower levels of CEFR B2 and are at or below the 25th percentile for graduate applicants.

If you have any questions about OEPP policies, ITA certification, or the use and interpretation of language proficiency test scores, please contact the OEPP at oepp@purdue.edu.

Cc: Linda Mason, Dean of the Graduate School
Tom Atkinson, Associate Dean of the Graduate School
Joel Ebarb, Associate Dean for Undergraduate Education and International Programs, College

THINK BROADLY. LEAD BOLDLY. COLLEGE OF LIBERAL ARTS
Young Hall | Room 810 | 155 S. Grant Street | West Lafayette, IN | 47907
765-494-9380 | purdue.edu/oepp

TOEFL iBT Subscale Scores, Cut Scores, and Common European Framework of Reference (CEFR) Descriptors

CEFR Level	TOEFL iBT Speaking	TOEFL iBT Writing	TOEFL iBT Reading	TOEFL iBT Listening
C2	-	-	-	-
	-	-	-	-
	-	-	-	-
	-	-	-	-
C1	30	30	30	30
	29	29	29	29
	28	28	28	28
	27	27	27	27
	26	26	26	26
	25	25	25	25
B2	24	24	24	24
	23	23	23	23
	22	22	22	22
	21	21	21	21
	20	20	20	20
B1	19	19	19	19
	18	18	18	18
	17	17	17	17
	16	16	16	16
A1 & A2	15	15	15	15
	14	14	14	14
	13	13	13	13
	12	12	12	12
	11	11	11	11
	10	10	10	10
	9	9	9	9
	8	8	8	8
	7	7	7	7
	6	6	6	6
	5	5	5	5
	4	4	4	4

Cut score Interpretations

Typical graduate admissions cut score of Purdue's aspirational peers, e.g., University of Michigan, Carnegie Mellon, MIT, UIUC (Engineering and Business)

25 is at the 75th percentile for graduate test takers.

OEPP recommended cut score for prospective ITAs: 100 total with no subscale score less than 22

22 is at the 50th percentile for graduate test takers.

Typical *undergraduate* admissions cut scores for Big Ten universities

20 is at the 40th percentile for graduate test takers.

Current Purdue graduate admissions cut scores

These scores are all below the 25th percentile for graduate test takers.

Common European Framework of Reference (CEFR) Descriptors

A **C2** level of English is essentially the level expected of a first language speaker. C2 allows for reading and writing of any type on any subject, nuanced expression of emotions and opinions, and active participation in any academic or professional setting. TOEFL is not designed to reliably measure C2.

C1 is the level at which a student can comfortably participate in all graduate activities, including teaching.

B2 measures the level required to participate independently in higher level language interaction. It is typically the level required to be able to follow academic level instruction and to participate in academic education, including both coursework and student life. However, B2 is an advanced intermediate level of language proficiency. Students entering graduate studies at B2 will benefit from language support, e.g., ENGL 620 or PLaCE short courses.

B1 is insufficient for full academic level participation in language activities. A student at this level could 'get by' in everyday situations independently. To be successful in communication in university settings, additional English language courses are required.

A1 and **A2** are insufficient levels for academic level participation.

REFERENCES

Alderson, J. C. (1984). Reading in a foreign language: a reading problem or a language problem? In Alderson, J. C. and Urquhart, A. H., editors, *Reading in a foreign language*, Essex: Longman.

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing, 26*(4), 621-631.

Al Fraidan, A. (2011). *Test-taking strategies of EFL learners on two vocabulary tests*. Germany: Lap Lambert Publications.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing, 9*(2), 101-119.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, *9*(4), 30-32.

Anderson, S. (2013). The importance of international students to America. *National Foundation for American Policy Brief,* 1-22.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66.

Axelson, E. R., & Madden, C. G. (1994). Discourse strategies for ITAs across instructional contexts In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 153–186). Alexandria, VA: TESOL.

Bachman, L. F., & Cohen, A. D. (2002). *Interfaces between Second Language Acquisition and Language Testing Research*. Beijing: Foreign Language Teaching and Research Press.

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Baker, B. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *19*(1), 63-83.

Bailey, K. M. (1984). *Foreign Teaching Assistants in US Universities. National Association for Foreign Student Affairs*. Washington, DC: Clearinghouse. Retrieved from https://eric.ed.gov/?id=ED249843.

Baker, B. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, *19*(1), 63-83.

Baker, B. A., Tsushima, R., & Wang, S. (2014). Investigating language assessment literacy: Collaboration between assessment specialists and Canadian university admissions officers. *Language Learning in Higher Education*, *4*(1), 137-157.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph No. 25). Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.Block, J. H. (ed.) (1971). *Mastery learning: Theory and practice*. New York: Holt, Rinehat and Winston.

Boltanski, L., & Thévenot, L. (2006). *On justification: Economies of worth* (Vol. 27). Princeton University Press, Princeton, NJ.

Boyles, P. (2005). Assessment literacy. In M. Rosenbusch (Ed.), *National assessment summit papers* (pp. 11–15). Ames, IA: Iowa State University.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, *33*(3), 307-318.

Burns, R. L. (1970). *Graduate Admissions and Fellowship Selection Policies and Procedures*. Princeton, NJ: Educational Testing Service.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series Report No. 20). Princeton, NJ: Educational Testing Service.

Carmichael, O. (1961). *Graduate education: A critique and a program*. New York, NY: Harper.

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. (2011). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19-27.

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

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

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, A. D. (1984). On taking language tests: what the students report. *Language Testing, 1*(1), 70–81.

Cohen, A. D. (2006). The Coming of Age of Research on Test-Taking Strategies, *Language Assessment Quarterly, 3*(4), 307-331.

Cohen, A. D. (2009). (Speaker). 11. TTS [12-minute internet video interview]. In G. Fulcher & R. Trasher (Eds.), *Language testing videos*. Retrieved from http://www.languagetesting.info/video/main.html#list. 26 June 2009.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (Monograph No. 33). Princeton, NJ: ETS.

Cotos, E., & Chung, Y. R. (2018). Domain Description: Validating the Interpretation of the TOEFL iBT® Speaking Scores for International Teaching Assistant Screening and Certification Purposes. *ETS Research Report Series*, *2018*(1), 1-24.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Crocker, L., & Algina, J. (1986). *Introduction to modern and classical test theory*. Fla: Holt Rinehart & Winston.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL* (TOEFL Monograph No. 26). Princeton, NJ: Educational Testing Service.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3): 327–347.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251-272.

Deygers, B., & Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36*(3), 347-368.

Dollerup, C., Glahn, E. and Rosenberg-Hansen, C. (1982). Reading strategies and test-solving techniques in an EFL-reading comprehension test - a preliminary report. *Journal of Applied Language Study 1*, 93-99.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series No. MS-8). Princeton, NJ: Educational Testing Service.

Douglas, D., & Hegelheimer, V. (2005). *Cognitive processes and use of knowledge in performing new TOEFL listening tasks* (2nd Interim Report to Educational Testing Service). Ames: Iowa State University.

Educational Testing Service. (2008). *Test and Score Data Summary for TOEFL [R] Internet-Based and Paper-Based Tests*. January 2008-December 2009 Test Data. ERIC Clearinghouse.

Educational Testing Service (2011). TOEFL iBT® research insight: TOEFL® program history. *Princeton, NJ: Educational Testing Service*.

Educational Testing Service. (2011). Reliability and comparability of TOEFL iBT scores. *TOEFL iBT Research Insight*, *1*(3), 1-8.

Elder, C. (2017). Language Assessment in Higher Education. In *Language Testing and Assessment (Encyclopedia of Language and Education*, pp. 271-286). Cham: Springer International Publishing.

Enright, M., & Tyson, E. (2011). Validity evidence supporting the interpretation and use of TOEFL iBT scores. *Princeton, NJ: TOEFL iBT Research Insight*.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113-132.

Ginther, A. (2003). International Teaching Assistant Testing: Policies and Methods. In Douglas, D. (Ed.), *English language testing in U.S. colleges and universities* (pp. 57–84). Washington, DC: NAFSA: Association of International Educators.

Ginther, A., & Elder, C. (2011, June 23-25). *Cutscores and Success* [conference presentation]. Language Testing Research Colloquium, Ann Arbor, MI, United States.

Ginther, A., & Elder, C. (2014). A comparative investigation into understandings and uses of the TOEFL iBT® test, the International English Language Testing Service (Academic) test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts. *ETS Research Report Series*, *2014*(2), 1-39.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271-295.

GlenMaye, L., & Oakes, M. (2002). Assessing suitability of MSW applicants through objective scoring of personal statements. *Journal of Social Work Education, 38*(1), 67-82.

Gu, M. (2018). An introduction to China's college English test (CET). *World Education News and Reviews*. Retrieved from https://wenr.wes.org/2018/08/an-introduction-to-chinas-college-english-test-cet

Gumport, P. J., Iannozzi, M., Shaman, S., & Zemsky, R. (1997). *Trends in United States higher education from massification to post massification.* Stanford, CA: National Center for Postsecondary Improvement, School of Education, Stanford University.

Haan, J. E. (2009). *ESL and Internationalization at Purdue University: A History and Analysis.* (Unpublished doctoral dissertation, Purdue University, West Lafayette, IN, United States). Retrieved from https://purdue.alma.exlibrisgroup.com/.

Haiyan, M., & Rilong, L. (2016). A Closer Look at Chinese EFL Learners' Test-Wiseness Strategies in Reading Test. *World Journal of Education, 6*(1), 68-74.

Hall, J. D., O'Connell, A. B., & Cook, J. G. (2017). Predictors of student productivity in biomedical graduate school applications. *PLoSOne,12*(1), e0169121.

Harmon, L. R. (1966). *Fourteen years of research on fellowship selection: A summary* (Vol. 1420). National Academies. Washington, DC.

Hollis, E. (1945). Forces That Have Shaped Doctoral Work. *Bulletin of the American Association of University Professors (1915-1955), 31*(3), 357-382. doi:10.2307/40220615.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, (25)* 328–402.

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

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kent, J. D., & McCarthy, T. M. (2016). *Holistic review in graduate admissions.* Washington, DC: Council of Graduate Schools.

King, G., Bruce, J. M., & Gilligan, M. (1993). The science of political science graduate admissions. *PS: Political Science and Politics, (26)*772–778.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' 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.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the graduate record examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement, (70*)340–352.

Landa, M. (1988). Training international students as teaching assistants. In J. A. Mestenhauser, G. Marty, & I. Steglitz (Eds.), *Culture, learning, and the disciplines: Theory and practice in cross-cultural orientation*. Washington, D.C.: NAFSA.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement, 17*(2), 28-30.

Lo Bianco, J. (2001). Policy literacy. *Language and education, 15*(2-3), 212-227.

Lowell, A. (1932). New Standards in Selecting Graduate Students. *Bulletin of the American Association of University Professors (1915-1955), 18*(3), 208-210. doi:10.2307/40218404.

Malone, M. E. (2017). Training in Language Assessment. In Language Testing and Assessment (*Encyclopedia of Language and Education*, pp. 225-239). Cham: Springer International Publishing.

Mamary, E. M., & Roe, K. M. (2004). Selecting for a diverse public health workforce – Community health education MPH program admissions at the California State University. *California Journal of Health Promotion, 2*(1), 22–28.

Marks, E. A. (2011). Ambiguous assessment: Critiquing the anthropology graduate admissions process. *Journal of Contemporary Anthropology, 2*(1), Article 4.

McPherron, P. (2017). Introduction: Why Study Globalization and Culture through English-Language Learning and Teaching in China?. In *Internationalizing Teaching, Localizing Learning* (pp. 1-39). Palgrave Macmillan, London.

Messick, S. (1987). Validity. *ETS Research Report Series*, 1987(2), i-208.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed. (pp. 13–103). New York: Macmillan.

Michel, R.S., Belur, V., Naemi, B. and Kell, H.J. (2019), Graduate Admissions Practices: A Targeted Review of the Literature. *ETS Research Report Series, 2019:* 1-18.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. viii−xx). Cambridge, UK: Cambridge University Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

National Center for Education Statistics (2019). *Postbaccalaureate Enrollment*. https://nces.ed.gov/programs/coe/pdf/Indicator_CHB/coe_chb_2019_05.pdf

National Foundation for American Policy. (2017). *The importance of international students to American Science and Engineering.* Arlington, VA: National Foundation for American Policy.

National Foundation for American Policy. (2013). *The importance of international students to America*. Arlington, VA: National Foundation for American Policy.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing (6)*199-215.

North, B. (2014a). *The CEFR in practice. English profile studies* (Vol. 4). Cambridge, UK: Cambridge University Press.

Okahana, H., & Zhou, E. (2019a). *Graduate enrollment and degrees: 2008 to 2018*. Washington, DC: Council of Graduate Schools.

Okahana, H., & Zhou, E. (2019b). *International graduate applications and enrollment: Fall 2018*. Washington, DC: Council of Graduate Schools.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*(3), 363-380.

Orfield, G. (2014). Realizing the promise of the civil rights revolution: Challenges and consequences for graduate education. *American Journal of Education, (120)*451–456.

Oppenheim, N. (1998, March). *Undergraduates' assessment of international teaching assistants' communicative competence.* Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Seattle, WA. Retrieved from ERIC Document Reproduction Service. (ED 423783).

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels*. Ed. James Carlson. New Jersey: Educational Testing Service.

Pierce, B. (1994). The test of English as a foreign language: developing items for reading comprehension. In: Hill, C., Parry, K. (Eds.), *From Testing to Assessment: English as an International Language*. Longman, New York, pp. 39-60.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing*, *30*(3), 381-402.

Pitcher, B. & Schrader, W. B. (1972). Indicators of College Quality as Predictors of Success in Graduate Schools of Business. *Admission Test for Graduate Study in Business, brief No. 6.* Princeton, NJ: Educational Testing Service.

Posselt, J. R. (2016). *Inside graduate admissions*. Harvard University Press: Cambridge, Massachusetts.

Powers, D., & Powers, A. (2015). The incremental contribution of TOEIC listening, reading, speaking, and writing tests to predicting performance on real-life English language tasks. *Language Testing, (32)*151–167.

Reilly, R. R. (1976). Factors in Graduate Student Performance. *American Educational Research Journal*, *13*(2), 125–138.

Rock, D. A. (1974). *The prediction of doctorate attainment in psychology, mathematics, and chemistry*. Princeton, NJ: GRE Board Research Rep.

Roever, C., & McNamara, T. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258.

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*(2), 159-183.

Rosenfeld, M., Leung, P., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. 21). Princeton, NJ: Educational Testing Service.

Sanyal, J. (2019). Use of English Language as a medium of instruction in imparting higher education at the post graduate level in India. *International Journal of English Learning & Teaching Skills, 2*(2).

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5–30.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309e327.

Schwager, I. T., Hülsheger, U. R., Bridgeman, B., & Lang, J. W. (2015). Graduate student selection: Graduate record examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European university. *International Journal of Selection and Assessment, (23)*71–79.

Spolsky, B. (1995). *Measured Words*. Oxford University Press, Oxford.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study. *American Psychologist, 52*(6), 630.

Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. *Handbook of formative assessment*, 233-250.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Report RR-08-34). Princeton, NJ: Educational Testing Service.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29,* 21-36.

Thomas, C., & Monoson, P. (1991). Issues related to the state-mandated English language proficiency requirements. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 382–392). Dubuque, IA: Kendall/Hunt.

Toulmin, S. (1958/2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

U.S. Bureau of Labor Statistics. (2019). *Employment, wages, and projected change in employment by typical entry-level education* (Employment in thousands). Retrieved from https://www.bls.gov/emp/tables/education-summary.htm.

Wagner, E. (2016). A study of the use of the TOEFL iBT® test speaking and listening scores for international teaching assistant screening. *ETS Research Report Series*, *2016*(1), 1-48.

Willingham, W. W. (1974). *Predicting success in graduate education. Science, 183*(4122), 273-278.

Wylie, E., & Tannenbaum, R. (2006). *TOEFL academic speaking test: Setting a cut score for international teaching assistants* (Research Memorandum No. RM-06-01). Princeton, NJ: Educational Testing Service.

Xi, X. (2008). Investigating the criterion-related validity of the TOEFL® speaking scores for ITA screening and setting standards for ITAs. *ETS Research Report Series*, 2008(1), i-65.

Xiao, Y. (2017). Formative assessment in a test-dominated context: How test practice can become more productive. *Language Assessment Quarterly, 14*(4), 295-311.

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, (58)* 149-162.

Yang, P. (2000). *Effects of Test-wiseness upon Performance on the Test of English as a Foreign Language* [Unpublished Dissertation]. Alberta, Canada: University of Alberta.

Yeom, S. & Jun, H. (2020). Young Korean EFL Learners' Reading and Test-Taking Strategies in a Paper and a Computer-Based Reading Comprehension Tests. *Language Assessment Quarterly, 17*(3), 282-299.

Zhang, L. (2009). Do foreign doctorate recipients displace US doctorate recipients at US universities?. In Ehrenberg, R. G. & Kuh, C. V., eds., *Doctoral Education and the Faculty of the Future*, 209-223. Itaca, NY: Cornell University Press.