5-1-2014

A THEMATIC ANALYSIS OF EXPERIENCES OF NON-NATIVE ENGLISH SPEAKING INTERNATIONAL GRADUATE STUDENTS WITH THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE

Annaliese Mayette

Annaliese M. Mayette
Candidate

Language Literacy and Sociocultural Studies
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Julia Scherba de Valenzuela, Chairperson

Jill Morford

J. Anne Calhoon

Jan Armstrong

A THEMATIC ANALYSIS OF EXPERIENCES OF NON-NATIVE ENGLISH SPEAKING INTERNATIONAL GRADUATE STUDENTS WITH THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE

by

ANNALIESE M. MAYETTE

B.A. Liberal Arts, University of Arizona, 1987
M.S. Experimental Psychology, UTEP, 1990

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor pf Philosophy
Educational Linguistics

The University of New Mexico

Albuquerque, New Mexico

May, 2014

Acknowledgements

I would like to thank the students who participated in this research. I am grateful for their gift to me of their stories. I hope I have warranted their trust in me to share their stories respectfully and honestly.

I thank my friends who have believed in me, and knew I could complete this dissertation even when I was not so certain. For food and coffee, for listening to me ramble on, for hugs and prayers, for stern talking-tos, for posting the Doctor Who original theme the day I passed my defense, and for so much more; I think you.

I thank the current and former members of Julia's doctoral writing group. Their comments and gentle criticisms have helped to make this a better document. I appreciate their help in all the forms it took, from editing drafts, to bringing snacks for my defense. I am also thankful to my committee members for their support over the past several years while I conducted this research and wrote this dissertation. Their thoughtful comments and suggestions contributed to the success of this research.

I am deeply grateful to Dr Julia Ann Scherba de Valenzuela for her patient direction, unwavering support, and encouragement. From my first week in this program through the completion of my degree I have known that I could always count on her support. Her advice and mentoring as I wrote and re-wrote the chapters of this dissertation have substantively contributed to the quality of the final document. I am a better student, theorist, and researcher due to her guidance and mentoring.

A THEMATIC ANALYSIS OF EXPERIENCES OF NON-NATIVE ENGLISH SPEAKING INTERNATIONAL GRADUATE STUDENTS WITH THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE

by

Annaliese M. Mayette

B.A. Liberal Arts, University of Arizona, 1987
M.S. Experimental Psychology, UTEP, 1990

Ph.D. Educational Linguistics, 2014

Abstract

First-person reports of perceptions and experiences of test takers is lacking in the literature. All stakeholders add to the understanding of the test. However, the test takers are the only stakeholders who have the experience of preparing for the test, taking the test, and living with the consequences of the test. This interview study reported on the experiences and perceptions of graduate students at one public university in the U.S. who have successfully taken the internet-based TOEFL. This research suggests that direct methods of eliciting opinions and experiences may be essential as participants described experiencing problems which they do not report to the ETS.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

George Bernard Shaw is credited with the quip that England and the United States are ‘two countries separated by a common language’. The extent to which speakers of English world-wide are “separated by a common language” is even greater in the 21st century as currently English is used or spoken by more non-native speakers than native speakers (Kachru, 1985, 1992). McCrum (2010) stated that an estimated one billion people speak English, most as a second language. As a global lingua franca, usage of English may be disambiguated from cultural and social ties to traditionally English-speaking peoples (Seidlhofer, 2001). While some (e.g., McCrum, 2010) have argued that this use of English is viewed as neutral, not carrying the negative association of the British or American imperialism that lead to the global spread of English, others (e.g., Templer, 2004) see it as a form of linguistic hegemony.

This global usage of English leads to varying norms among different populations of speakers and users of English (Kachru, 1992). The concept of language proficiency is frequently misunderstood (Goh, 2004), and this diverse pattern of usage further complicates defining (Nelson, 1992), and assessing (Lowenberg, 2002) language proficiency in English. In universities in Canada and the United States the solution to the complex problem of assessing academic English language proficiency in second language (L2) speakers is most often accomplished by the use of the Test of English as a Foreign Language (TOEFL) (Zareva, 2005). This test, a product of the Educational Testing Service (ETS), has changed over the decades of its existence reflecting changes in

language theory, test design theory and practice, technology, and the needs of various stakeholders (Biber et al., 2004).

Background of the Problem

TOEFL history and development. The development of the TOEFL began in 1961 when a varied group of stakeholders (from government, assessment organizations, and universities) met to discuss the creation of a test that was inexpensive in both time or cost to administer, tested all essential elements of English, and had demonstrable objectivity, reliability, and validity (Spolsky, 1990). In short, they wanted it all. The need for a psychometrically valid, objective, and standardized assessment that "should be based on specifications of actual needs" (Spolsky, 1990, p. 107) was agreed upon by the attendees. What the essential elements of English were, or at least what essential elements needed to be assessed, was perhaps a point of disagreement. In the end, those elements of language that were more easily tested via then modern psychometric assessment techniques (such as language structure, vocabulary, reading comprehension) were included in the original test, while those elements that proved harder to test objectively (such as oral comprehension, oral production, written production) were not included in the earliest versions of the TOEFL (Spolsky, 1990).

The TOEFL went through several iterations, changing format from paper based to computer based, to internet-based (Zareva, 2005) as technological changes allowed. Assessment of additional aspects of academic English usage were added over time, including the Test of Written English (TWE) and the Test of Spoken English (TSE) (Stansfield, 1986). The latest version of the TOEFL, the internet-based (INB) TOEFL, incorporated all of these subtests into one multipart assessment (Zareva, 2005). The INB

TOEFL also added tasks that required integration of multiple aspects of language proficiency (Enright, 2004).

Changes to the TOEFL were driven by technical advances, advances in testing theory and practice, and also by requests from the stakeholders. In a discussion of the TOEFL 2000 framework, Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000) specified that their constituencies "primarily include score users in North American colleges and university undergraduate and graduate admissions community, applied linguists, language testers, and second language teachers" (p. 3). With a few notable exceptions (e.g., Rosenfeld, Leung, & Oltman, 2003; Stricker & Attali, 2010) student test takers have rarely been directly included in ETS research, suggesting that the ETS may view their input as less important than that of other stakeholder groups. A few recent studies have addressed test taker experiences and perceptions of the TOEFL (e.g., (He & Shi, 2008; Huang, 2006; Stricker & Attali, 2010; Yu, 2007).

The TOEFL is the assessment most commonly used to assess international student's mastery of academic English by US and Canadian colleges and universities (Zareva, 2005). It is, however, not the only commonly available test of academic English proficiency. In addition to the TOEFL, the Michigan Test of English Language proficiency (MTELP), and the International English Language Testing System (IELTS) are also commonly used assessments of English language proficiency (Templer, 2004) that international students at the post-secondary level may opt to take instead of the TOEFL at some universities.

These post-secondary level tests, such as the TOEFL, also serve non-academic purposes. In addition to use by educational institutions, other groups such as corporations

(Templer, 2004), high schools, embassies, licensing agencies, governments, professional boards, and language schools (Jamieson et al., 2000) also use such tests. However, the ETS produces the Test of English for International Communication (TOEIC) and the Secondary Level English Proficiency test (SLEP) that would be more appropriate for some of these alternative uses. The TOEIC is designed to assess English as used in corporate and other non-academic environments (Wilson, 1989). The SLEP was designed to assess the English language proficiency of students at the secondary education level for whom English was not a native or first language (ETS). I believe that the use of the TOEFL by these other groups, and for purposes other than assessing academic English as used at colleges and universities in Canada and the United States, must be questioned given both the stated purpose of the TOEFL and the availability of other English language proficiency exams.

International student enrollment. I believe that issues of English language proficiency assessment in higher education will only increase in importance as the number of international students at American institutions of higher education increases. According to the Open Doors survey (Institute of International Education, 2010), during the 2009-2010 academic year the total enrollment of international students studying at colleges and universities in the United States was over 690,000. According to this survey the two countries with the largest enrollments in US institutions of higher education continue to be India and China, with each showing increases over the previous year (30% and 2% respectively). The reputation of American academic credentials and degrees, and the student's English language proficiency were two of the leading reasons cited by

international students in their choice to come to the US for higher education (Obst & Forester, 2006).

At the University of New Mexico (UNM), the institution where this study was conducted, there were 1,040 international students enrolled in Fall 2011. In Fall 2010 there were 970 international students (University of New Mexico Division of Enrollment Management, 2011). Overall, in the United States there are more international undergraduate students than international graduate or professional students (Institute of International Education, 2010). Of the international students in a degree seeking status at UNM, 185 were undergraduates, and 591 were in graduate or professional programs (UNM Division of Enrollment Management, 2011). Many of the international students were exchange students, and thus were in non-degree status while at UNM. This group accounted for 264 of the international students in Fall 2011 (University of New Mexico Division of Enrollment Management, 2011).

As is the norm for most English language medium colleges and universities in the United States, international student applicants to UNM from countries where English is not among the official or national languages must take an English language proficiency exam. The TOEFL was the most common test taken by international students at UNM (personal communication with Anne Barnes, UNM international admissions officer, January 2011). The IELTS is also accepted as proof of English language proficiency. Students who submit other entrance exam scores (e.g., ACT, SAT, GRE) may be exempt from submitting scores for an English language proficiency test depending on their score on the language component of these other tests.

English and Global Education

In the beginning of the 21st century, English is arguably the global lingua franca (Mauranen, 2003; Seidlhofer, 2001, 2005), especially in the fields of commerce and education (Fulcher, 2007; Mauranen, 2003). English is "considered to be an asset that can lead to success in the 21st century job market" (Tsai & Tsou, 2009, p. 319). Crystal (2000) argued that this contextualized use of English as a language of broader communication has contributed to its expansion of use. Kachru (1985) described three groups of English speakers. His first group includes those for whom English is the traditional cultural or national language. Use of English initially spread beyond the cultures and nations of traditional usage via geographic contact and colonization (Kachru, 1985). More recently, English has spread to cultures and nations beyond the sphere of direct contact or colonization (Kachru, 1985; Kachru 1992). People in this third group had many reasons for acquiring facility in English. For many individuals speaking English opens access to education in traditionally English speaking countries (Fulcher, 2007).

Globally, and particularly in the traditionally English speaking countries, education is increasingly seen as a commodity or product of the market economy (Fulcher, 2008; Gillan, Damachis, & McGuire, 2003; Grace, 1989; Halic, Greenberg, & Paulus, 2009; Hursh, 2007; Longhurst, 1996; McMurtry, 1991; McPerren, 2007; Naidoo, 2007; Noble, 2003). This commodification of education has occurred at the primary/secondary level (Hursh, 2007), and the tertiary level (Grace, 1989; Fulcher, 2007; Halic et al., 2009; Longhurst, 1996; Noble, 2003). This change in the conceptualization of education has taken place in the UK (Grace, 1989; Fulcher, 2007;

Longhurst, 1996), in the US (Halic et al., 2009; Nobel, 2003; McPherren, 2007), and in Australia/New Zealand (Gillan et al., 2003; Selvarajah, 2006).

The shift towards commodification of higher education has been associated with internationalization of higher education (Gillen et al., 2003; Naidoo, 2007; Halic et al., 2009). While many have seen this shift in the conception of education unfavorably, other authors have identified some positive benefits (Naidoo, 2007; Selvarajah, 2006; Wildavsky, 2010). For example, some countries, such as the People's Republic of China (PRC) have a population growth rate that has far outstripped the development of educational infrastructure at the tertiary level (Naidoo, 2007). In other countries, the higher education infrastructure has available capacity greater than is needed for its own population (Selvarajah, 2006). This suggests that there may be short term benefits to both the countries that send, and the countries that receive higher education students. However, the long term effects of this commodification and internationalization of education at the tertiary level on students who travel out of their home country for college, the colleges they attend, and the sending and receiving countries remain unknown.

In the US the number of residents aged 18 peaked in 2009, and is expected to decline in the following years (Hussar & Bailey, 2008). As the number of these traditionally “college aged” US residents decreases the number of international students whom US institutions may accommodate may increase due to surplus capacity (Selvarajah, 2006). I believe that in the short-term accepting more international students may therefore benefit these institutions. However, in the longer-term there may be unexpected costs, complications, and challenges for the US institutions, and the

international students who come to them. I believe that as international student registration at US colleges and universities increases, and as awareness of the economic value of these students increases, the need for and interest in research on academic English language proficiency assessments should also increase.

The benefits gained by international students who attend colleges and universities in traditionally English speaking countries vary, but must be perceived as significant given the number of international students who apply each year to colleges and universities in traditionally English speaking countries. I believe that the investments in time and money that are required to attain English language proficiency levels needed for study at universities where English is the language of instruction might suggest perceived benefits of study in traditionally English speaking countries for these international students. Additionally, in most cases, students have to study for, take, and pass some assessment of English language proficiency to be considered for admission. In North America the most common English language proficiency assessment for entry into tertiary education is the Test of English as a Foreign Language, the TOEFL (Zareva, 2005). Therefore, I believe that research on the INB TOEFL is both timely and important.

Statement of the Problem

As stated previously, I find the minimal inclusion of the test takers to be a gap in the research. Rea-Dickins describes stakeholders as “those who make decisions, and those who are affected by those decisions” (Rea-Dickins, 1997, p. 304). Hamp-Lyons (2000) stated that “of all the stakeholders in testing events, test takers surely have the highest stake of all” (p. 581). However, test takers are rarely included in design processes of or research on the test that they will take (Hamp-Lyons & Lynch, 1998). These test

takers are the ones who pay for the test, study for the test, take the test, and live with the consequences of their test results. Therefore, I believe that in any reasonable analysis of a test, test takers' input should be considered, including their experiences with it, and its effect on them.

Purpose of the Study

First-person reports of perceptions and experiences of people who have taken the INB TOEFL are lacking in the literature. I argue their perspective is important as the test takers are the group of stakeholders with the greatest personal experience. They are also the group most directly affected by the process and use of the test. Therefore, the purpose of this exploratory study is to report experiences and perceptions with the INB TOEFL from international graduate students at UNM who have taken the test.

Questions to be Addressed

The two questions that I addressed in this study were:

1. What are the perceived experiences of non-native English speaking international graduate students with the internet-based version of the TOEFL?
2. What are these students' perceptions of the applicability of the internet-based TOEFL in light of their subsequent experiences with academic English?

Conceptual Assumptions, Researcher Stance, and Operational Definitions

Conceptual assumptions. I take a descriptive, functionalist approach to language, rejecting linguistic prescriptivism and structuralism. From the descriptive perspective all variants of a language are valid, and no variety (e.g., dialect or register) of a language is inherently superior to another. From a functionalist approach language is a purposive communicative process situated within social contexts. These conceptions of

language are important to this study as they inform my understanding of language proficiency.

Researcher stance. Although I reject positivism and its associated assumption of the researcher as expert, I embrace empiricism and its emphasis on methodological precision and structure. I also believe that context is important. I want to understand experiences from that more holistic perspective, including the context in which they occur. Further, I am most interested in the personal and individual, rather than social or structural aspects of situations. As such, I find phenomenology, with its focus on the lived experience of a specific situation or condition, to be a satisfying tradition and process through which to ask and answer questions of experiences and perceptions. I believe that the stories of individuals, while personal and unique, can shed light on experiences or conditions that many people share. For these reasons, I find phenomenological interviewing to be a technique particularly well suited for addressing issues of personal experiences.

With regard to the questions addressed in this research; I am not an international student and I have not taken any version of the TOEFL. As a student I did not find standardized tests to be particularly anxiety producing or disturbing. I found many assessments to be rather game-like; not that they were particularly fun, but rather that they followed observable or perceivable rules, and probabilities. I have been interested in assessment, both traditional and alternative for many years.

Over the past two decades I have worked in the fields of social epidemiology and higher educational research as a statistician, programmer/analyst, and data manager. Coming from a mostly quantitative background, my interest in the internet-based TOEFL

was initially related primarily to the predictive validity of the test. As I immersed myself in the literature my focus changed. My interest in the specific topic of this research is based on my conception of fairness and justice within education in general and assessment in specific, as well as interactions with international students who have taken the INB TOEFL.

I assume that test designers have paid little attention to the input of test takers, parents, teachers, or other professionals (therapists, diagnosticians, etc) who administer the test they design. I assume that the corporations that produce large-scale standardized tests are at least as interested in their reputation and bottom line as they are in producing good, valid, and meaningful tests.

Operational definitions. For the purposes of this dissertation I used the following definitions:

- English - any of the many global Englishes, and varieties and registers there of.
- High stakes test - an assessment with great potential impact on the test taker, such as high school exit exams, college entrance exams, citizenship exams.
- Language proficiency - communicative competence in a given language including expressive and receptive skills.
- Stakeholder (in a test) - any person or organization with an interest in the implementation, use or interpretation of a test, particularly those who are directly affected such as students/test takers, teachers and parents.
- Standardized test – a commercial test that is administered in a set way to all test takers, without regard to the social or cultural background of the test takers.

Rationale and Theoretical Framework

The theoretical framework that guides this study is naturalistic and qualitative (Lincoln & Guba, 1985). The qualitative tradition that will inform the methods and analysis of this study is phenomenology. I did not precisely follow any one researcher's methods of phenomenology or phenomenological interviewing, but was influenced by the descriptions of phenomenological research as presented by several researchers (c.f., Creswell, 1998; Moustakas, 1994; Seidman, 2006; Smith, Flowers, & Larkin, 2009; Van Manen, 1990). The primary research method that I used in this study was qualitative interviewing.

Importance of the Study

The importance of this study is that it presents the reported experiences and perceptions of test takers. This group of test stakeholders is infrequently included in published research (Hamp-Lyons & Lynch, 1998). This study addressed this gap in the literature. I believe that there is power in speaking one's truth; participants in this study may have experienced this through participation in this study.

Scope and Delimitations of the Study

The purpose of this research was to address the perceived experiences of non-native English speaking international graduate students with the internet-based version of the TOEFL, and their perceptions of the applicability of the INB TOEFL in light of their subsequent experiences with academic English at a university in the US. In this research I exclusively addressed the INB TOEFL. I did not address any other test of English language proficiency, other formats of the TOEFL, or associations between the experiences of the test takers and their TOEFL test scores. In this research I included only international graduate students at the University of New Mexico. I did not address other

populations of individuals taking English language proficiency assessments. This study was qualitative and included a thematic analysis of the interview texts. I did not make any comparisons of the individual, nation-based or language-based differences in experiences with the INB version of the TOEFL. Additionally, in this research I did not address registers or uses of English other than academic English. Although the research questions addressed the experiences of international graduate students with the INB TOEFL and their subsequent academic English usage, I know that what I actually received from my participants were self-reports of their perceptions and recollections of those experiences.

Chapter 2

Review of Literature

With my research questions in mind, I have described the relevant literature in the following areas: (a) the role of standardized testing in US education, (b) the TOEFL and the ETS, (c) a description of the internet-based TOEFL, and (d) the role of "stakeholders" in high stakes assessment in education.

The Role of Standardized Testing in US Education

The No Child Left Behind (NCLB) act of 2000 made several changes to US public education at the Pre-Kindergarten through 12th grade level (P-12). One of the most wide ranging and controversial pieces of this act was the requirement for greater reliance upon high-stakes standardized tests (Garrison, 2009). This should have come as no surprise to politicians, as previously the increased testing in US public elementary and secondary schools in response to the Nation at Risk report of 1983 had also been controversial (DeMerle, 2006). Standardized tests have been used in US public schools and for admissions decisions for US colleges and universities since the mid-1800s (Garrison, 2009). While the purposes and stakes associated with these tests have varied over time and by location, Haladyna, Haas, and Allison (1998) stated that "achievement tests always have been used by the public to evaluate educational progress" (p. 262), and that "US schools have used tests to weed out students and eliminate them from further education opportunities" (p. 262). Full histories of standardized testing in American schools have been published (e.g., Clarke, Madaus, Horn, & Ramos, 2000; Haertel & Herman, 2005; Pelligrino, 2004). It is not my intention to replicate those here. Rather, I provide a summary of those aspects most related to this research.

History of Standardized Testing in the US

The history of standardized testing in US schools reflects social change, and the expansion of educational opportunity to social, economic, gender, and ethnic groups not previously included in public education (Clarke et al., 2000; Garrison, 2009; Haladyna, et al., 1998; Lemann, 2000). Often, decision-makers viewed these students as less capable than the previous limited population of students (Lemann, 2000). The motivation for establishing standardized educational testing came from a deficit theory of American educational institutions; the assessors were measuring failure not success (Clarke et al., 2000; Garrison, 2009). Recent policies that have led to increases in standardized testing are also often seen by some as based on the assumption of the failure of the American educational system (Garrison, 2009). These perspectives and assumptions affect the selection and implementation of assessments, I believe.

The development of standardized testing is related to the expansion of educational opportunity. “The first documented achievement tests were administered in the period 1840 to 1875, when American educators changed their focus from educating the elite to educating the masses” (Haladyna, et al., 1998, p. 262). Some researchers (Clarke et al., 2000; Garrison, 2009) argue that the general public and policy makers show a desire to measure the failure of the educational system as educational opportunity expands beyond just the elites to the majority of the population as it did in the middle to late 19th century in the US. Early forms of standardized tests were developed to measure this ‘failure’ (Clarke et al., 2000), and to prove the need for school reform (Office of Technology Assessment, 1992). One of the first large scale implementations of standardized tests in the United States was in the Boston public schools and coincided with the move to

educate more of the populace (Garrison, 2009). The widespread belief in the powers and objectivity of science helped further the development of standardized tests (Clarke et al., 2000). Many of these tests were highly biased against cultural and linguistic minorities, and reinforced biases against different cultural groups, particularly new immigrants (Haladyna et al., 1989). At a national level, the Army Alpha, an early variant of an intelligence test, was an instrument designed to help the military determine which jobs to assign new recruits (Lemann, 2000). It came into use in 1917. Its use heralded in an age of testing in the US (Smyth, 2008).

Lemann (2000) sorts the developers of standardized test who worked between the two world wars into four groups: the progressives, IQ test designers, the standards imposers, and the education expanders. Three of the groups (the progressives, IQ tests designers, and standards imposers) were from, or working in elite colleges and universities (Lemann, 2000). Although only one group was directly interested in development of IQ tests (following from Thorndike), all of these three groups based their assessments on intelligence tests or the Army Alpha, that itself was based on intelligence tests (Lemann, 2000). There were large differences between these groups, but they all embraced some form of elitism, so for my purposes I will consider them one group. One form of elitism that they embraced, meritocracy (Lemann, 2000), can be described as “a particular type of vertical classification that is centered on competition as a basis for ranking and thus status and power” (Garrison, 2009, p. 12). Educational tests that were developed within this paradigm include the SAT and GRE (Lemann, 2000).

Lemann’s (2000) fourth group, the education expanders, who he describes as ultimately losing to the others, came out of a public university in Iowa, and was led by E. F.

Lindquist. Lindquist was not seeking a meritocracy, instead he "wanted to educate more students not fewer and to use tests to further that goal" (Lemann, 2000, p. 25). Having worked on a test meant to identify the best students in the state early in his career, he spent most of his career working on achievement tests meant for all students such as the Iowa Test of Basic Skills and the ACT (Lemann, 2000). In contrast to the SAT which aimed to assess a student's aptitude, "E.F. Lindquist's creation of the ACT in 1959 as a competitor to the SAT, intended as a measure of achievement rather than ability" (Atkinson & Geiser, 2009). Lindquist believed that achievement tests should have diagnostic components and be educationally useful (Atkinson & Geiser, 2009). In addition to tests for public school students, and college entrance, Lindquist was also involved in the creation of the General Education Development (GED) test, an alternative credential equivalent to a high school diploma (Batmale, 1948). The GED was designed to assist returning veterans' efforts to further their education, and use their veteran's education benefits, by providing an alternate to a high school diploma which could be used for college entrance (Batmale, 1948).

In contrast, the other test developers were not interested in educating more students. Ben Wood, a student of Thorndike and one of the early writers of standardized tests, believed that too many people were getting into colleges, and that "testing would purge the educational system of its pervasive idiocy" (Lemann, 2000, p. 35). Wood went on to develop the Graduate Record Exam in 1935 (Lemann, 2000), a test which some researchers (Schonemann & Heene, 2009) claim continues to be biased against people from non-dominant culture groups. Wood and Lindquist therefore represent opposite

ends of what could be seen as a spectrum of educational inclusion, with Lindquist wanting to educate all and Wood wanting to limit education to those most ‘gifted’.

Meritocracy. Some proponents of meritocracy believed this notion was supported by Thomas Jefferson (Lemann, 2000). In letters to John Adams, Jefferson describes a natural aristocracy who were the right people to lead and make decisions for the newly birthed nation (Garrison, 2009). Adams' response was one of disgust and general opposition to the creation of any sort of aristocracy in the US (Lemann, 2000). Although individual purposes may have differed, the net effect is that these early standardized tests were developed and used in a manner congruent with Jefferson’s concept of natural aristocracy, or as it would later be called, meritocracy. The creation of this meritocracy, or rule by the most gifted or able, would require distinguishing between the more and less capable. Early standardized tests were biased against cultural and ethnic groups (Demerle, 2006), and the less intelligent (as indicated by their tests) who were believed to “inevitably gravitate towards immoral and criminal behavior” (Garrison, 2009, p. 12). For many of the test developers the biases may have been unintentional and perhaps even unnoticed by these men based on the social and cultural norms of the time. However, according to Lemann (2000) some test developers openly embraced eugenics (selective breeding of humans) and therefore may not have been blind to the biases and effects of the tests they developed.

The development of the SAT came from the desire to find this natural aristocracy. The test was created to help the deans of elite schools (originally Harvard) find those deemed worthy of an elite education based on presumed merit so they could be offered scholarships and the advantages of elite private education (Lemann, 2000). Non-

scholarship students were not required to take the SAT, as their ‘merit’ could be determined based on their secondary school records. In time the use of the SAT expanded to public schools and all applicants for admission in private schools. This growth was related to the expansion of applications to higher education after World War two, and particularly to the expansion of applications of ethnic and cultural minorities (Lemann, 2000; Garrison, 2009). Although the SAT developers began with the objective of finding those of high merit regardless of background (Lemann, 2000), Hamp-Lyons argued that the “meritocracy they were designing with their 'objective' tools was shaped in their own image” (Hamp-Lyons, 2000, p. 583).

With the advent of scoring machines for multiple choice tests, standardized testing really took off in the 1950s. This technological advancement influenced both the type and number of tests administered (Haladyna et al., 1998). At the same time, the tradition of 'educational reform' based on the notion of failure continued (Garrison, 2009). National, state and local mandates for testing increased with each wave of educational reform (Clarke et al., 2000; Garrison, 2009).

Increases in use of standardized testing. Successive waves of education reform, and the subsequent increases in educational testing, happened throughout the second half of the 20th century (Garrison, 2009). In the 1950s the post-Sputnik race to catch up with the Russians was a driving motivation in education (Clarke et al., 2000). In the 1980s the National Commission on Educational Excellence published ‘The Nation at Risk’ based upon the idea that American schools were failing, as the chair of the commission later revealed (Clarke et al., 2000). This commission was not hired to “objectively examine the condition of US public education” (Garrison, 2009, p. 106), but to “document the bad

things…about public schools" (Garrison, 2009, p. 106). Reaction to 'The Nation at Risk' lead to the Educate America Act of 1993 (Clarke et al., 2000). The No Child Left Behind Act of 2001 continued the traditions of a belief in the failure of American schools, and of requiring additional standardized testing (Hertel & Herman, 2005). Clarke et al. stated that "in fact most educational reforms now rely heavily on testing to serve a multitude of purposes" (Clarke et al., 2005, p. 159). Even prior to the implementation of the No Child Left Behind Act of 2001 that has again increased testing requirements, Kohn stated that "children are tested to an extent that is unprecedented in our history and unparalleled anywhere else in the world" (Kohn, 2000, p. 2). The primary consumers and beneficiaries of testing were policy makers (Pelligrino, 2004).

Another major beneficiary of increased standardized testing was the test publishing corporations (Bracey, 2005). Clarke et al. (2005) estimated that US elementary and secondary students took a combined 400 million tests per year in 2005. With the subsequent increased testing requirements for NCLB compliance (Garrison, 2009), this number was likely much higher in the 2010-2011 academic year. Clearly the testing agencies have strong motivation to lobby for increased usage of standardized testing. The potential for the testing agencies to become influential in education policy was seen by some of the early test designers (Lemann, 2000). Brigham, one of the developers of the Army Alpha, objected to mass testing later in life, as "what worried him most, because of his long experience with incautious testers (including himself in his younger days), was that any organization that owned the rights to a particular test would inevitably become more interested in promoting it than in honestly researching its effectiveness" (Lemann, 2000, p. 40).

Pushback. With each wave of increased standardized testing there has been pushback from researchers, parents, and educators (Demerle, 2006). Researchers from Margaret Mead (Mead, 1926, 1927), Walter Lippmann in the 1930s (Lemann, 2000) to Stephen Krashen (Krashen, 2011) have objected loudly and often to the use of standardized testing asserting that they are biased, and not educationally useful. In the late 1990's while surveys showed that most parents supported the use of standardized tests in public schools (Haladyna et al., 1998) a protest and boycott movement was gaining support (Demerle, 2006). At that same time, educators and professional associations also opposed increased use of and higher stakes uses of standardized testing (Kohn, 2000).

By the late 1990s the parents' groups protesting standardized testing were apparently considered newsworthy, as stories ran in major media outlets about parents from Massachusetts to California keeping their children home on 'test day' (DeMerle, 2006; Kohn, 2000). Lawsuits were brought against school districts that required standardized testing (Demerle, 2006). Demerle stated that protesting and boycotting standardized testing was seen across the US, but least in the Southeast US. According to Hamp-Lyons (2000) pushback against testing is a phenomenon found mainly in Australia and the US. These parents were increasingly organized (Demerle, 2006) and getting the attention of policy makers. Perhaps in part as a backlash against this boycotting of testing, the NCLB initially required 95% of students in each school and in each specified sub-population take the test in order for a school to pass adequate yearly progress (AYP) (Sunderman, 2006). There have been several changes to the interpretations and

implementations of this rule over time, including implementation of exceptions (Sunderman, 2006).

Grassroots movements against increased use of standardized testing are supported by online resources. Groups such as FairTest, that "draws in teachers and testing professionals, but is primarily driven by, and identified with, lobbies of parents and students" (Hamp-Lyons, 2000, p. 579), and writers/bloggers who are opposed to standardized testing such, as Susan Ohanian, are easily found on the internet. In 2012 the U.S. Department of Education issued NCLB waivers for 10 states (United States Department of Education, 2012) and stated that they expected to issue more waivers (United States Department of Education, 2012).

With increasing emphasis on standardized testing and increasing stakes related to the results of standardized testing it should not be surprising that cheating on these tests occurs. Based on data prior to implementation of NCLB testing requirements, Jacob and Levitt (2003) stated that "serious cases of teacher or administrator cheating on standardized tests occur in a minimum of 4-5 percent of elementary school classrooms annually" (p. 843). In the years since then, there have been several reports of teacher cheating on standardized tests (Beckett, 2011), and other 'fabrications' of data (Koyama, 2011) required under NCLB.

Accurate or inaccurate, test scores are king in the current US educational climate. Primary and secondary level students are assessed with NCLB required standardized testing. Teachers and students are evaluated based on the results of these standardized tests. Students entering or continuing higher education also face standardized testing that

may include the SAT, ACT, TOEFL, GRE, MCAT, LSAT, and others depending on the student and the level or program to which the student is applying.

The TOEFL and the ETS

By the 1980s "the fastest growing test by far was the TOEFL" (Lemann, 2000, p. 242). The internet-based TOEFL (INB TOEFL) is the current version of the TOEFL, an assessment of academic English language proficiency produced and administered by the ETS (Zareva, 2005). It is the most common high stakes standardized test used by colleges and universities in English speaking countries, especially in Canada and the United States, to assess non-native English speaking international student's English language proficiency (Zareva, 2005). The internet-based version has been used since 2006 (Zareva, 2005). Previous versions include the paper-based version of the TOEFL, and the computer adaptive version of the TOEFL (Educational Testing Service, 2003). In this section I will summarize the history of the TOEFL, and present a description of the INB TOEFL, including a summary of its development, related research, and concerns about the INB TOEFL.

The TOEFL has been produced and administered by the ETS since 1964 (Spolsky, 1990, 1995). The TOEFL is a test of academic English as used in colleges and universities in traditionally English speaking countries, particularly Canada and the United States (Zareva, 2005). Following from this, the intended use is as an "assessment of university level English language skills" (Biber et al., 2004, p. 1).

The ETS. The ETS describes itself as a non-profit organization that "continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the

organization's products and services" (Cohen & Upton, 2006, p. ii). The ETS produces the TOEFL, the GRE, and other large scale standardized tests (see http://www.ets.org/tests_products). Although the ETS is a non-profit company, it does charge test takers for their products (TOEFL, SAT, TEOIC, etc.). In 2011 test takers paid between $150 and $225 to take the TOEFL, where the exact amount varied depending on the country in which the test was administered (http://www.ets.org/toefl/ibt/about/fees/). As a comparison, in 2011 the GRE general test cost between $160 and $190 depending on location, the GRE subject tests cost between $140 and $160 depending on location (see http://www.ets.org/gre/ revised_general/about/ fees/), and the SAT cost registrants $47 in 2011 with an available fee waiver for those who showed financial need (see http://sat.collegeboard.org/register/sat-fees).

In the early 1960s when the TOEFL was first designed, psychometrics and objectivity were considered to be two of the main requirements for good tests (Spolsky, 1995). At that time, the primary psychometric quality of concern for most researchers was reliability (Xi, 2008). By the 1980s validity was the primary psychometric quality of concern to most researchers (Xi, 2008). For a review of reliability and validity research on versions of the TOEFL other than the internet-based version see Chapelle, Grabe, and Burns (1997) and Hale, Stansfield, and Duran (1984). The psychometric qualities of the test remain important to the ETS, as evidenced by the volume of reports on this topic that they produce annually (see http://www.ets.org/toefl/research).

Changes to the TOEFL. There have been many changes to the TOEFL since it was first administered in 1964 as I detailed in Chapter One. By the time the ETS began developing the INB TOEFL they described the TOEFL as developing within a

"framework that takes into account models of communicative competence" (Cohen & Upton, 2006, p. ii). The current version includes assessment of expressive English skills (Zhang, 2008), and includes performance based items (Zareva, 2005). This contrasts with the early versions of the TOEFL test that assessed mostly receptive English skills via multiple choice questions (Spolsky, 1990).

Some of the changes came about due to strong pressure by universities, English language teachers, and other stakeholders. Although the ETS maintained "that it was simply not possible to test the writing ability of hundreds of thousands of candidates by means of a composition: it was impractical, and any how the results would be unreliable" (Hughes, 2003, p. 6), in 1986 the ETS began use of the TWE. Hughes (2003) stated that "the principle reason given for this change was pressure from English language teachers" (Hughes, 2003, p. 6). While this was a win for this group of stakeholders, it was at best a moderated win, as "scorers of the TOEFL Test of Written English have just one and one half minutes for each scoring of a composition" (Hughes, 2003, p. 95).

Internet-Based TOEFL

The internet-based TOEFL (INB TOEFL) is the current version of the TOEFL (Alderson, 2009). Implementation of the INB TOEFL began in 2005 in limited centers, with full implementation at all testing centers the next year (Zareva, 2005). The INB TOEFL has four sections, Reading, Writing, Listening and Speaking (Alderson, 2009). Two of these sections, Writing and Speaking, have both independent and integrated tasks (Enright, 2004). The test uses both performance tasks and multiple choice questions (Enright, 2004). "The reason behind the test revision is the realization that to succeed in an academic environment in which English is the language of instruction, students need

not only to understand English, but to communicate effectively" (Zareva, 2005, p. 46). The addition of performance based tasks and the integrated tasks aligns the INB TOEFL more closely with the sorts of English language tasks that test takers will perform in US and Canadian colleges (Alderson, 2009). This also brings the test more in line with current conceptions of language proficiency tests (Hughes, 2003).

As the name suggests, the INB TOEFL is completed online. It is administered 30 to 40 times a year at more than 4,300 testing centers world-wide (Alderson, 2009). Test takers are allowed up to four hours to complete the test, with a maximum of 30 minutes for each of the parts of the test (Alderson, 2009). Test takers are allowed to take notes during the listening section (Alderson, 2009) which is new for this test. Total scores range from a minimum of 0 to a maximum of 120, with ranges of 0 to 30 on each of the four sections (Alderson, 2009).

Texts used. The texts used in the Reading and Listening sections differ from those used in previous versions of the TOEFL Reading section (Enright, 2004). The texts used in the TOEFL Computer Based Test (CBT) were described as "like entries in an encyclopedia" (Anon, 2003, p. 117). In response to criticism such as this and in response to stakeholders, the ETS supported studies of academic English texts as used in US colleges and universities (Biber et al., 2004). Biber et al. (2004) stated that there were "few large scale empirical investigations of academic registers, and virtually no such investigations of spoken registers" (Biber et al., 2004, p. v). To answer this gap in the literature, Biber et al. (2004) performed a corpus study of actual academic English texts. They included both written and spoken texts. Their analysis of these texts showed that some previous assumptions about academic English were not accurate (Biber et al.,

2004). Although they collected multiple real world texts, none were used in the actual INB TOEFL as they were all considered to be too specific to their given domains (Enright, 2004). The texts used were constructed based on the characteristics Biber et al. (2004) found in the corpus texts (Enright, 2004).

Receptive language. The INB TOEFL tests both reading and listening academic English skills. A major consideration for receptive language tests is that the "texts employed in the test reflect salient features of the texts the test takers will encounter in the target situations as well as demonstrating the comparability of the cognitive processing demands of accompanying test tasks with target reading activities" (Green, Unaldi, & Weir, 2010, p. 191). The texts designed for the reading and listening sections were created based on Biber et al.'s 2004 analysis of academic English texts (Enright, 2004). One of the salient features of the texts is the complexity of the text. Text complexity is described variously, but some common components are vocabulary, syntax, and inference/reference (Green et al., 2010). As in the previous versions, the INB TOEFL includes comprehension and inference questions that are multiple choice format (Cohen & Upton, 2007). A new component of the reading section, reading to learn, requires the students to interact more with the texts (Gomez, Noah, Schedl, Wright, & Yolkut, 2007). The reading to learn tasks were assumed by the test designers to be more difficult than the comprehension and inference questions (Cohen & Upton, 2007). These assumptions informed the scale anchoring of the reading test (Gomez et al., 2007). However, in a study of test takers' verbal protocols Cohen and Upton (2007) found that the reading to learn tasks "were among the easiest" (p. 224) for test takers, and that the "newer formats were not more difficult than the more traditional formats" (p. 234).

Interactions between the specific domain of the texts and test taker's prior knowledge with regard to previous TOEFL reading and writing sections have been shown or suggested (He & Shi, 2008; Nissan, DeVincenzi, & Tang, 1996). In regards to listening tasks some researchers found that there are also domain knowledge effects (Kostin, 2004; Nissan et al., 1996; Sadighi & Zare, 2006) While the prior knowledge effects may not be large (Kostin, 2004; Nissan et al., 1996), they did affect test takers scores on the listening test (Sadighi & Zare, 2006). Further, Sadighi and Zare (2006) found a significant effect of topic priming on listening test score.

The listening section is new for the INB TOEFL (Kostin, 2004), and reflects change to the TOEFL based on the needs of stakeholders, as a need for a test that reflects academic English lecture participation (listening and note taking) has been reported in the literature (Huang, 2004, 2006). Sawaki and Nissan (2009) stated that this section was "designed to assess academic listening ability in the context of academic lectures and conversations that take place in various situations on campus" (p. 1). Although, associations between the TOEFL listening sub-score and other measures of academic listening skills have been shown (Sawaki & Nissan, 2009), this section of the INB TOEFL is open to criticisms related to context based purpose for the test takers. Kostin describes the texts in the listening section as occurring within "sparse linguistic context" (Kostin, 2004, p. 2). The listening texts are presented out of broader context, therefore the test taker has no a priori knowledge of what will be important in the spoken texts (Kostin, 2004). Based on this criticism, I believe that it is likely that this listening task may be influenced by context effects and prior domain knowledge even more than reading tasks

as the spoken text is only presented once, but the test taker is allowed to read the written tasks more than once before going on to the questions about it.

Expressive language. The writing and speaking sections of the INB TOEFL assess a test taker's ability to express themselves in academic English. These sections derive from the Test of Written English (TWE) and the Test of Spoken English (TSE) that were previously optional additional tests of academic English offered by the ETS (Educational Testing Service, 2011). These sections require the test taker to perform individual tasks and also to perform tasks that integrate multiple modalities (Alderson, 2009). ETS stated that "integrated tasks require test takers to combine their English-language skills, as is typically done when communicating in an academic setting" (Educational Testing Service, 2011, p. 5). In addition to the scores on these sections, as of 2009 the ETS "allows score users to listen to a 60-second portion of an applicants scored speaking response to one of the TOEFL iBT integrated speaking tasks" (Educational Testing Service, 2011, p. 5). As with the previous section, prior domain knowledge of test takers has been an issue raised in research (Kostin, 2004).

Adding these performance based tasks increases the complexity of scoring the test as "performance tasks are time consuming to administer and to score, and this imposes severe practical constraints limiting the number of tasks administered and ratings obtained in large-scale standardized assessment contexts" (Enright & Quinlan, 2010, p. 318). However, even with limited examples Enright and Quinlan (2010) claimed that "these timed writing exercises are sufficient to provide evidence of examinees' basic writing skills" (p. 319). Considerations of time and cost influenced the decision to implement an automated rating program as "automated scoring of writing has the

potential to dramatically reduce the costs associated with large-scale writing assessments" (Weigle, 2010, p. 335).

INB TOEFL writing samples are scored by both human raters and an automated program, e-rater (Enright & Quinlan, 2010). This use of automated scoring "has led to controversy" (Weigle, 2010, p. 335), as "acceptance of human scoring is high despite known limits" (Enright & Quinlan, 2010, p. 318), but automated electronic scoring "meets with less acceptance" (Enright & Quinlan, 2010, p. 318). Texts are scored at least twice, once by a human rater and once by the e-rater (Enright & Quinlan, 2010). The text is rated only by two human raters if the essay is determined by the first human rater to be off topic, or the e-rater finds too many grammatical errors. Additionally, if there is a discrepancy between the human and e-rater scores, a second human rater reads and scores the written text (Enright & Quinlan, 2010).

The score generated by the e-rater is mostly a product of text length and grammatical features (Enright & Quinlan, 2010). However "some e-rater feature scores are associated with human holistic scores even when length is taken into account" (Enright & Quinlan, 2010, p. 326). This may be in part due to the e-rater model being weighted to "optimize prediction of human scores" (Enright & Quinlan, 2010, p. 330). Overall, in defending the use of e-rater Enright and Quinlan (2010) stated that "correlations between a variety of criteria of writing skills and the scores on the TOEFL independent essays were mostly in the range of 0.30 to 0.40. These correlations were only slightly higher for human scorers than for e-rater scores" (Enright & Quinlan, 2010, p. 328). This does not seem like a strong argument for the e-rater to me, but rather as a moderate argument against the validity of the independent writing task of the TOEFL.

Beyond any arguments about cost, and time, and correlations of scores, is the question of the purpose of writing. If one takes the perspective that "writing is primarily a means of communicating between people, not a collection of measurable features of text" (Weigle, 2010, p. 349) then opposition to the e-rater cannot be overcome by demonstrating correlations with human raters' scores. On the other hand, even as the ETS "affirms that writing is fundamentally a social act" (Enright & Quinlan, 2010, p. 330) they may be constrained by limits of cost and time. Not all test takers may be opposed to the e-rater. Acceptance of or objection to any scoring rubric or methodologies may be influenced by test taker group (Yu, 2007).

Cultural and linguistic backgrounds of test takers are likely to influence their spoken texts (Carey, Mannell, & Dunn, 2010; Chalhoub-Deville & Wigglesworth, 2005). Culture and first language related variation can affect scores on the speaking section of the TOEFL (Carey et al., 2010; Chalhoub-Deville & Wigglesworth, 2005). Carey et al. (2010) found that raters who were second language English speakers gave higher scores for spoken texts overall. This was true for both test takers who were from their linguistic or cultural group as well as for those who were not in their linguistic or cultural group. Chalhoub-Deville and Wigglesworth (2005) found that US raters gave higher scores than raters from other countries (Canada, Australia, and the UK), while raters from the UK gave lower scores than raters from other countries. Although the effect sizes were small, all differences between US and UK raters were significant (Chalhoub-Deville & Wigglesworth, 2005). Some researchers have found that self-reported facility with spoken English correlated well with spoken language scores (Powers, Kim, Yu, Weng, & VanWinkle, 2009). While none of these studies used the INB TOEFL speaking section

(Carey et al., 2010 used recordings from the IELTS speaking test, Chalhoub-Deville and Wigglesworth, 2005 used the TSE, and Powers et al., 2009 addressed the spoken section of the TOEIC), I believe that rater effects may also be found for the INB TOEFL speaking section.

Some researchers (Iwashita, Brown, McNamara, & O'Hagan, 2008; Xi, 2007) have addressed the INB TOEFL speaking section in their research. Xi (2007) looked at the viability of providing analytic scores using the component scores on the speaking test. However, Xi found that component scores of the speaking test were too highly correlated to be reliable as independent measures, and therefore they could not be used individually to provide additional information on a test taker's specific speaking skills. Iwashita et al. (2008) looked at the distinctiveness of level scores on the speaking section. While they did find level effects, they "were not as great as might have been predicted" (Iwashita et al., 2008, p. 41). They found that vocabulary and fluency had the greatest influence on test taker's speaking score, and that both grammar and pronunciation also contributed. They describe pronunciation's role as "a sort of first level hurdle" (Iwashita et al., 2008, p. 44).

Factor structure. The factor structure of a test does not necessarily follow the format structure of a test. Factor analysis can reveal the relationships among tasks on a test that can differ from the intended structure of the test. The intended structure of the INB TOEFL is one overall score (higher order factor) and four sub-scores (first order factors). In an analysis of a pre-release version of the INB TOEFL Stricker, Rock and Lee (2005) found two first order factors but no higher order factor. The speaking sub-test contributed to one factor, while the combination of reading writing and listening together

contributed to the other factor. Sawaki, Stricker and Oranje (2009) performed a confirmatory factor analysis using the INB TOEFL in the form in which it was released in 2005. They found a structure closer to the expected, with one higher order factor and four first order factors, speaking, reading, writing, and listening each contributed to separate factors. They also found that the speaking factor differed from the others, showing the lowest association with the higher order factor, and lower associations with other factors. This may in part be related to concerns raised by some researchers (see for example Carey, Mannell, & Dunn, 2010; Chalhoub-Deville & Wigglesworth, 2005), as described in the previous section.

The Role of Stakeholders

As discussed in Chapter One, stakeholders play an important role in the design, evaluation, and re-design of tests. Their input may lead to changes in a test. Some stakeholders' voices are more influential than others, as "the power of some stakeholders is far greater than that of others" (Hamp-Lyons, 2000, p. 588). Teachers', parents', and students' or test takers' voices are considered to be among the least influential (Rea-Dickens, 1997). In my opinion, their input should be among the most important. If one takes issues of consequential validity and test fairness seriously, then the input of those most affected by the test is essential. Beyond and above test design and implantation issues are human issues; students/test takers, parents, teachers, and others who administer tests are all affected by tests and testing (Fulcher & Davidson, 2007). For this reason I believe that their voices should be considered. However, as I will show in the following sections these voices are not often attended to by test development corporations.

Consequential validity, fairness, and justice. Researchers and writers in the field of ethics in testing vary in their use of terms related to ethics of testing (McNamara & Ryan, 2011). Their organization of these terms and concepts varies also (McNamara & Ryan, 2011). Some described testing ethics principally in terms of fairness (Hamp-Lyons & Lynch, 1998; Kane, 2010; Kunnan, 2004; Xi, 2010), justice (McNamara & Ryan, 2011), responsibility (Hamp-Lyons, 2000), or consequential validity (Messick, 1989). McNamara and Ryan (2011) stated that "the principal writers on the ethics of language testing… discuss issues of fairness and justice more or less interchangeable" (p. 165). This focus on the consequences and fairness of tests stems in large part from Messick's (1989) arguments for and descriptions of consequential validity of tests (Popham, 1997).

Xi (2010) links test fairness to test validity. Although some (Kunnan, 2010) disagree with this specific argument, the general thesis that test fairness is part of, or related to the validity of a test continues a line of theory going back at least to Messick (1989). Messick (1989) proposed that the consequences of an assessment should be considered as part of the validity argument of the assessment thereby bringing the concept of consequential validity to the area of test validity. He stated that the tests must be considered not only "as the means to the intended end, but in light of other ends it may inadvertently serve, and in consideration of the place of the intended end in the pluralistic framework of social choices" (Messick, 1989, p. 85). Some researchers, like Popham (1997), rejected the idea of consequential validity. Others (Hamp-Lyons, 2000; Kane, 2010; Kunnan, 2004; McNamara, 2001) worked in, worked on, and further developed the concept of fairness and consequences of testing. McNamara highlighted the importance of the concept of fairness and consequences when he stated that "the fundamentally

social character of language assessment challenges us to rethink our priorities and responsibilities in language testing research" (McNamara, 2001, p. 333).

McNamara and Ryan (2011) stated that "a concern for fairness is fundamental to theories of validity in language testing" (p. 162). If one accepts that as a given, then without fairness a test fails to be valid. This makes the need for a definition and testable operationalization of fairness important to test developers and researchers. However, Kunnan (2004) stated that in regards to test fairness "there is no coherent framework that can be used for evaluating test and testing practice" (p. 27). His framework, the test fairness framework, includes five elements: "validity, absence of bias, access, administration, and social consequences" (Kunnan, 2004, p. 27). In this framework, he separated consequential validity from other forms of validity (content, construct, criterion, and reliability), naming it social consequences. The test fairness framework defined fairness as a concept superordinate to validity.

Xi's (2010) model was presented as "an approach to guide practitioners on fairness research and practices" (p. 147). It presented steps and assumptions that researchers and test designers could follow to check for test fairness based on an argument structure described by "six inferential steps and the mechanisms under which they can be organized conceptually to link an observation in a test to score-based interpretations and uses" (Xi, 2010, p.156). Xi (2010) argued against viewing fairness as an isolated concept or as a superordinate concept, but rather described fairness as an aspect of validity, specifically as "comparable validity across groups within the population of interest" (p.147). Xi (2010) added to the previous works by the addition of an argument structure that would assist test designers in assessing test fairness. This

model allowed for specific testable hypotheses related to fairness. Several writers on test ethics have challenged Xi's assumptions or definitions. For example, Kunnan (2010) disagreed with the inclusion of the argument structure as unneeded and not helpful, and also disagreed with the conception of fairness as an aspect of validity. Davies (2010) said of Xi's argument that "Xi convinces me with the strength of her validity argument. What I find puzzling is whether this has anything to do with fairness" (p. 176). He (Davies, 2010) argued that "validity as I understand it does itself apply to all comparable groups; why do then we need to appeal to fairness? Like justice, validity guarantees that an ability is being appropriately tested for a relevant population" (p. 175). McNamara and Ryan (2011) critiqued Xi's model of test fairness as they argued that it "elides the complexity of the highly contested values dimensions of tests, which need to be addresses in a different way, through direct political and ethical argumentation, discussions which will necessarily be open-ended given they are arguments about values" (p.167).

McNamara and Ryan (2011) distinguished between test fairness and justice in testing. They described fairness as a form of validity in that "validity theory has incorporated an expanded understanding of the social dimension of tests, which are now more widely recognized as social instruments serving social goals" (p.162). In their model, justice includes "considerations of the consequential bias of test score interpretation and use but also and particularly, the social and political values implicit in test construction" (p.167). Thus in their model, some aspects of 'fairness', as described by other authors, are a part of validity, while other aspects are beyond the scope of validity. In their model, McManara and Ryan (2011) call these other aspects 'justice'.

These researchers and theorists approach the issue of ethics in tests, test design, and test use from different angles and with differing theoretical or philosophical perspectives as are evident in their models or frameworks, the most well-known of which I described above. Allowing for all of these differences, all of the above cited researchers and theorists allow that some aspects of consequences, fairness, or justice are relevant in considerations of test ethics.

Stakeholder perceptions and experiences. Stakeholder is in itself a value laden term, with historical use coming from legal and business origins (Ohanian, 2000). Some researchers like Ohanian (2000) decline to use the term due to what she described as inequality and economic implications. Definitions for stakeholder vary by researcher for those who provide explicit in text definitions (Bachman & Palmer, 1996; Fulcher & Davidson, 2007; Ohanian, 2000; Rea-Dickins, 1997). The range of stakeholders may vary from individuals as in Bachman and Palmer's (1996) statement that

> a variety of individuals will be affected by and thus have an interest, or 'stake' in the use of a given test in any particular situation. Stakeholders that are directly affected include the test takers and test users, or decision makers (p. 31),

to large social groups as revealed in Fulcher and Davidson's (2007) observation that "in large-scale high-stakes tests the society of a country or region may have a stake in how a test is used" (p. 376). Another description allowed that stakeholders are "those who make decisions and those who are affected by the decisions" (Rea-Dickins, 1997, p. 304), with recognition that the more conventionally powerful are making the decisions and the less conventionally powerful are being affected by those decisions. Involving more stakeholder groups, particularly the less powerful would lead to the "democratization of

assessment processes" (Rea-Dickins, 1997, p. 311), however while test taker input is important "their views are among the most difficult to make sense of" (Rea-Dickins, 1997, p. 307). While I agree that stakeholders are "all individuals or organizations with an interest in the use of or impact of the test" (Fulcher & Davidson, 2003, p. 376) I will focus on research related to students/test takers stakeholders in the following.

While "reliance on 'top down' psychometric approaches in arguments for validity, reliability, and fairness" (Fairbairn & Fox, 2009, p. 16) remains more common than the "bottom up test taker feedback" (Fairbairn & Fox, 2009, p. 16), input from test takers is being elicited in some limited contexts and for certain purposes. For example researchers may collect data on the duration and type of test taker preparation for a test (Ren, Bryan, Min, & Wei, 2007), the process and procedures that a test taker employs in the process of completing a test (e.g., Cohen & Upton, 2007), the emotional and psychological effects of testing on a test taker (Triplett & Barksdale, 2005), the interaction of some characteristic of a test taker such as confidence or perceived ability with test outcome (Huang, 2006; Stankov & Lee, 2008), test taker preferences for formats or modalities of tests (Brown & Hirschfeld, 2008; Pino-Silva, 2007; Struyven, Dochy, & Janssens, 2002; Tsai & Tsou, 2009), test takers' perceptions of the validity of tests (Brown & Hirschfeld, 2008; He & Shi, 2008; Powers et al., 2009; Struyven et al., 2002; Tsai & Tsou, 2009; Yu, 2007), and test takers' experiences and attitudes about test processes as a whole (Cohen & Upton, 2007; Stricker & Attali, 2010; Stricker, Wilder, & Rock, 2004). Any of these may be compared to the score or outcome of the test. In the following I will focus on research on the preferences, perceptions, and experiences of adult test takers on tests in general, and on the TOEFL in particular.

Students' perceptions of assessment. A student's perceptions of assessment in general and of a given test in particular can have an effect on test outcomes. In a study involving elementary aged students, Brown and Hirschfeld (2008) found that students' basic conception of the role and nature of testing, along with how the meaning and purpose of a specific test is presented to them, both have calculable effects upon test outcome. Students who "conceive of assessment as a means of taking responsibility for their learning … demonstrate increased educational outcomes" (p. 3). When tests are presented to students as accountability measures for teachers, or schools, rather than for individual students, test "achievement is likely to go down" (Brown & Hirschfeld, 2008, p. 13). Tsai and Tsou (2009), in a study of university students, found that most students in their study were unhappy about the addition of a new standardized language proficiency assessment and believed that the standardized language proficiency assessment was "insufficient to reflect what was learned and taught" (p. 319). In their study, Tsai and Tsou (2009) found that students who had higher self-reported language skills and who reported feeling lower levels of stress and pressure over the test and test outcome had less negative opinions about the test. Triplett and Barksdale (2005), in a study involving American elementary school students, found that the day after administration of a standardized test, that many students were "anxious and angry" (p. 255) about several aspects of the test, test process, and test experience.

Test takers preferences for different types of test formats and modalities have also been studied, and this bottom up feedback can provide surprising and conflicting data. In a meta analysis of hundreds of previously published studies, Struyven et al. (2002) found that test takers viewed alternative assessments as less fair than traditional assessments.

However, they found that “many students perceived traditional assessment tasks as arbitrary and irrelevant” (p. 5). Similarly, Selvarajah (2006), in a study of international and native university students in New Zealand, found that although cultural differences existed, traditional individual assessments were seen as the fairest type of assessments by all cultural groups in the study. These traditional assessments were also seen as the least preferred assessments (Selvarajah, 2006). Selvarajah (2006) did find culture based differences in test format preferences, with international students from Asian countries preferring group assignments and assessments, while native New Zealand students did not prefer this type of assignment.

Some research has found that computer based tests may cause difficulty for some test takers (Dooey, 2008; Maycock & Green, 2005), or lead to lower scores (Manalo & Wolfe, 2000). Manalo and Wolfe (2000), in a study that compared comparable writing samples from paper based and computer based versions of the TOEFL, found that students’ hand written texts are scored higher than texts composed on a computer. However, Pino-Silva (2008), in a study of a new computerized language proficiency assessment, found that “test taker perceptions of the computerized test appear to be positive” (p. 148). One of the complaints of Pino-Silva’s participants was that test takers could not return to previous questions. This complaint was also found by Cohen and Upton (2007) in their study of the TOEFL test takers’ test taking strategies.

Although the inclusion of test takers’ perspectives are important for the democratization of assessment (Rea-Dickins, 1997) some researchers are equivocal on the validity of participants’ self-reported skills, abilities, or experiences (Huang, 2006). However, self-report provides data available from no other source, such as individual’s

perceptions. Powers et al. (2009), and Huang (2006) investigated test taker perceptions of their abilities, or confidence in those abilities. Yu (2007) presented an unexpected twist on test takers' opinions on the inclusion of their own voices, suggesting that in some cases test takers prefer an expert's scoring template over a student/test taker influenced template, even when their scores are lower with use of the expert template.

Powers et al. (2009) found that test takers' self-report of their perceptions of their language proficiency were positively correlated with their TOEIC test scores and test score components. This suggests that test taker self-report is reliable. Huang (2006) found that Chinese students (all of whom had scored high enough on the TOEFL to enter an academic program at an English language of instruction university in North America) reported low confidence in their English language proficiency. Specifically, "their self-ratings showed that listening, speaking, writing, pronunciation, and vocabulary are their weak areas" (p. 224). Huang adds the caveat that these findings are based on self-report and therefore may not "represent the real situation" (p. 225), suggesting a view that the takers' reality is not as relevant as some hypothetical objective reality.

Yu (2007) found that a small group of Chinese student test takers, when given the option of having their written summaries of test texts scored based on a popular (student based) template or a template created by experts, chose the expert template even though they scored higher when the popular template was used. For this specific group of students, the value of voice and democratization seemed to be of less importance than respect for the expert, and a perception of correctness as equated as similarity to an expert.

Some recent research has addressed test takers' perspectives on the TOEFL (Cohen & Upton, 2007; He & Shi, 2008; Stricker & Attali, 2010; Stricker et al., 2004). He and Shi (2008) found that their student participants felt that the TOEFL writing test was easier than a writing test that they were required to take and pass once they came to attend college in Canada. Their participants also expressed that the Canadian test was more culturally biased than the TOEFL. Like Huang's (2006) participants, He and Shi's participants had scored beyond the local minimum on the TOEFL, but still described difficulties with requirements once they began study at an English language of instruction institution.

Cohen and Upton (2007) and Stricker and Attali (2010) each conducted research on the INB TOEFL with support from the ETS. Both pairs of researchers expressed the importance of test taker feedback to test design and use. Cohen and Upton (2007) analyzed verbal protocols recorded while participants were taking the reading portion of the INB TOEFL. They looked at task, processing, and reader purposes, stating that "it is important to have a good insight into what it is people who take reading comprehension tests do in order to complete them" (p. 210). As such they were interested in both test taking and reading strategies and "the analysis paid close attention to whether the reported processes for responding to a given item were consistent with the aims of the test constructors and hence indicated that the item was testing what it purported to" (p. 223). They found that, unexpectedly, reading strategies were the same across the test, and that the parts of the test that were thought by the test developers to be more difficult were in fact identified by the participants as the easiest. Despite all of the research and development efforts, "the new formats were not more difficult than the traditional

formats" (p. 234). I believe that this highlights the importance of test taker feedback for the content and construct validity of a test.

Stricker and Attali's (2010) results were also unexpected. The previous research into the attitudes of TOEFL test takers had showed generally positive attitudes about the Computer-Base Test from all national groups included in the study (Stricker et al., 2004). In their 2010 study, Stricker and Attali found that on average participants from three of the four included national groups had positive to neutral attitudes about the INB TOEFL, and one had neutral to negative attitudes. These researchers suggested that current test takers may be more comfortable complaining about that test than test takers were previously, suggesting that the data from these two seemingly similar studies may therefore not be directly comparable. As Stricker and Attali's (2010) study participants scored above average on the test, it is possible that attitudes of all test takers may be even less positive than their study reported. These researchers embarked on their research based on an assumption that the value of data from test takers was primary for test validity, as "acceptance by test takers, test users, and the public is essential to the continued viability of the TOEFL" (Stricker & Attali 2001, p. 1). They end their discussion with a call for more studies of test taker attitudes, and also for a more "fine grained analysis of the test takers' attitudes about the TOEFL" (Stricker & Attali, 2010, p. 15).

In conclusion, as the literature I have reviewed suggests that the development of academic or ability testing in the US is often in the service of the dominant, powerful elites. In service to this, the testing corporations have become powerful themselves. The needs of some stakeholders have been privileged while the needs and even the input of

others have been marginalized or ignored. It is important to me that all voices are heard and all input considered in the development of assessments. This is important in my opinion not only for justice, but also for the validity of the assessments. No one knows the experience of taking a test like the test takers themselves. As Cohen and Upton (2007) found, input from test takers can provide a new and different perspective on an assessment, including the difficulty of the sections and the types of skills and knowledges employed in answering test questions. This suggests that professional, expert test designers who follow all standard procedures for test design can err in their assumptions about the subject matter or in how test takers will interact with the assessment.

Chapter 3

Methods

Overview

This study was designed to answer the following questions:

1. What are the perceived experiences of non-native English speaking international graduate students with the internet-based version of the TOEFL?

2. What are these students' perceptions of the applicability of the internet-based TOEFL in light of their subsequent experiences with academic English?

To answer my research questions I interviewed eight non-native English speaking international graduate students in their second or subsequent semester at the University of New Mexico who have taken INB TOEFL. I conducted this research within a phenomenological qualitative framework. I employed phenomenological interviewing as my primary methodology.

The phenomenological interview process may involve a long interview (Moustakas, 1994) before which there may be some social conversation and a request for the participant to focus on the activity or event of interest (Moustakas, 1994). I began with a free association task and then led directly into the interview. Information shared during the free association task was included in the thematic analysis along with information shared in the interviews. Although primarily associated with clinical practice, free association tasks have been employed in a variety of research domains including marketing (Koll, von Wallpach, & Kreuzer, 2010), information science (Jung, Pawlowski, & Wiley-Patton, 2009), linguistic or corpora research (Lahlou, 1996; Viks-Freibergs & Freibergs, 1976; Wettler, Rapp, & Sedlmeier, 2005), and with other more

usual psychological tests for intelligence or ability testing (McClatchy & Cooper, 1924). Free association tasks access declarative knowledge (Koll et al., 2010), and are a good "first approach... that can then be completed with other more classical methods" (Lahlou, 1996, p. 279). I included the free association task as I hoped it would help to set a focus on the research topic in addition to providing data of a different sort than the interview (declarative in contrast to the more personal or episodic from the interview).

Research Design

This research study employed a qualitative, phenomenological, interview design. I analyzed participant interview transcripts via thematic analysis informed by phenomenological assumptions. My methodology was based primarily on Moustakas's methodology (Moustakas, 1994) involving one long open ended interview. The questions in the interview focused on the participants' experience of a specific phenomenon, here the experience of taking the internet-based TOEFL.

Description of Methodology

To address my research questions I employed a free association task, an interview protocol, and a member check protocol in this order for all research participants.

The free association task included terms related to the focus of the study (see Appendix A). The interview protocol contained queries related to participants' experiences of the TOEFL assessment and process, and their experiences with academic English language use subsequent to taking the TOEFL (see Appendix B for the interview protocol). The member check protocol contained queries related to the participants' responses to the themes that I identified in the interview (see Appendix C for the member check protocol).

The member check consisted of two parts. I contacted all participants and asked them to meet individually with me to discuss the themes I found in my analysis and request their input on whether I made correct assumptions as to their meaning for quotes from each of them that I included as exemplars of each theme, as per Appendix C. Additionally, when needed I requested a conversation in person, by phone, or email to clarify comments or topics from the interview, or to address other issues that came up in the process of transcription or during data analysis.

Selection of Participants

Eight English speaking international graduate students who were in their second or subsequent semester at the University of New Mexico and who had taken and passed the INB TOEFL were participants in this study. The participants were from a convenience sample with potential participants identified through fliers sent to international student organizations, and via snowball recruitment wherein colleagues shared fliers with students, coworkers, and friends). Although I used all these recruitment methods, what ultimately worked was snowball recruitment and references by friends. I was successful in my attempt to get the same number of male and female students, and also students from a variety of countries. These students were either master's or Ph.D. students. Students whose research focus, course work, or previous experience included language teaching, language theory, or related discipline and those who had a degree from an English speaking country, were excluded from participation. As I did not get more volunteers than my planned maximum number of participants, no selection process was used. Should more than 12 people who met these criteria have volunteered to participate I would have selected participants in order to balance gender and maximize

geographic region of origin. See Appendix D for the participant recruitment flier. Participants were volunteers and they were not paid for their participation in this study.

Data Collection and Recording

Data collected included the participants' responses to the free association task, their responses to the interview protocols, and information obtained during member checking. I digitally recorded and took notes during the free association task and the interview. Digital recordings of the free association task were used to assure that all responses were noted correctly during the task. The response from the free association task and the interview were the data for the thematic analysis.

Design and Procedure

I distributed participant recruitment fliers to professors, friends, and international graduate student organizations. When individuals interested in participating in this study contacted me I reminded them of the criteria for participation and asked them to confirm that they met the criteria. After confirming that the volunteers met criteria and answering any questions that they had, those who agreed to participate met me at a mutually agreed upon time and location for further discussion of the study, the consent process and, if they agreed to participate, to conduct the free association task and interview. The participants received and signed the informed consent form prior to beginning the study. See Appendix F for the informed consent form. All participants were free to terminate participation at any time during the study. I contacted participants after the interview to clarify any confusion I had and for more information if needed. Once I had analyzed the transcripts I sent emails requesting another meeting with the participants to share with them the themes that I found and to elicit their input on these themes. Two participants

agreed to meet for this member check. The free association task and interview together were of approximately 45 minutes to one hour and 45 minutes in duration. The member checks varied in duration from approximately 30 minute to one hour.

With participant permission, I digitally recorded all interviews. I saved the electronic files of these interviews to a USB drive and encrypted via PGP. I stored the USB drive in a locked file cabinet at my office. After I verified the accuracy of my transcription I deleted all digital voice files.

Data Processing and Analysis

I identified and coded themes from the transcripts using Dedoose online qualitative analysis software and based upon an interpretive phenomenological model of analysis. Throughout all stages of this research project I discussed my biases and assumptions with a 'critical friend' as a way of bracketing my pre-existing and ongoing assumptions. This process helped me to keep focused on the participants' experiences and voice.

In the coding stage I read through the transcripts multiple times. The first readings were to get a general understanding of the texts and to identify questions I had of the texts. During subsequent iterations of readings I identified meaning units in each transcript and coded them. After this I followed an iterative process of reviewing the codes for similarity and then combining codes into groups. I merged codes that were very similar and combined codes that related to common higher level constructs. In this fashion I build themes from the codes thusly identified. Later readings of the transcripts were primarily to clarify that the original texts (included in the themes) did in fact

support the associated themes with which I had identified them. This process was iterative and I returned to earlier steps in the process as needed.

Throughout this process of analysis I discussed and shared details about the data, analysis, and process with my advisor, “critical friend”, and colleagues in my advisor’s doctoral student research group. The purpose of this sharing was to check my biases and assure that the analysis represented the emic voice (Headland, 1990), and to understand this data through the different perspectives that these colleagues kindly shared with me.

Confidentiality and storage of data. I assigned pseudonyms to all interview participants and referred to the participants only by these pseudonyms in my transcriptions, my writing, and in discussions of this research, both in formal settings and with my group of 'critical friends'. I kept one hardcopy of the link with the pseudonyms and actual names that was stored in a locked file cabinet at my office. The hard copy of this link was destroyed after the transcripts and general characteristics of each participant were uploaded to Dedoose. I also stored the original signed informed consent forms in a locked file cabinet at my office. I encrypted (via PGP key) all electronic files of the digital recordings of the interviews. No one else had access to the files of the digital recordings of the interviews. Any personal information shared by the participants that was not pertinent to this research was not transcribed. In transcriptions I disguised information about specific location in country of origin (region or city) and previous institutions attended. See appendix E for transcription practices that I employed. Transcriptions of the interviews were shared with my advisor and my group of ‘critical friends’ as needed throughout the process of this research.

The confidentiality of persons who have suggested the names of possible participants was also protected, including from the participants themselves. Additionally, I did not inform anyone who shared or posted the recruitment fliers if any of those with whom they shared the fliers participated in this research project. Individuals sharing the recruitment fliers were not considered research participants in this project because no data was being collected from them. Therefore, informed consent was not be necessary for the nominators.

I destroyed all digital recordings of the interviews once they were transcribed and after those transcriptions had been checked for accuracy. I also destroyed the single hard copy of the mapping of actual names to pseudonyms once the transcripts and descriptors were uploaded to Dedoose.

Chapter 4

Results

Research does not happen in hypothetical space; its context informs participants, researchers, and analytical findings. Similarly, participants, researchers, and environment interact to create the contextual space in which the research takes place. Therefore, I will describe the context of this research briefly. This research was conducted at a large Southwestern university. The eight participants in this study all took the internet-based TOEFL (INB TOEFL) at non-US test sites, all passed the INB TOEFL, and all were subsequently admitted into graduate programs at this university (see Table 1 for basic demographics of the participants). Genders were equally represented, as was degree level (Masters Degree, or Doctoral Degree) of graduate program (see Figure 1 for gender and academic program level).

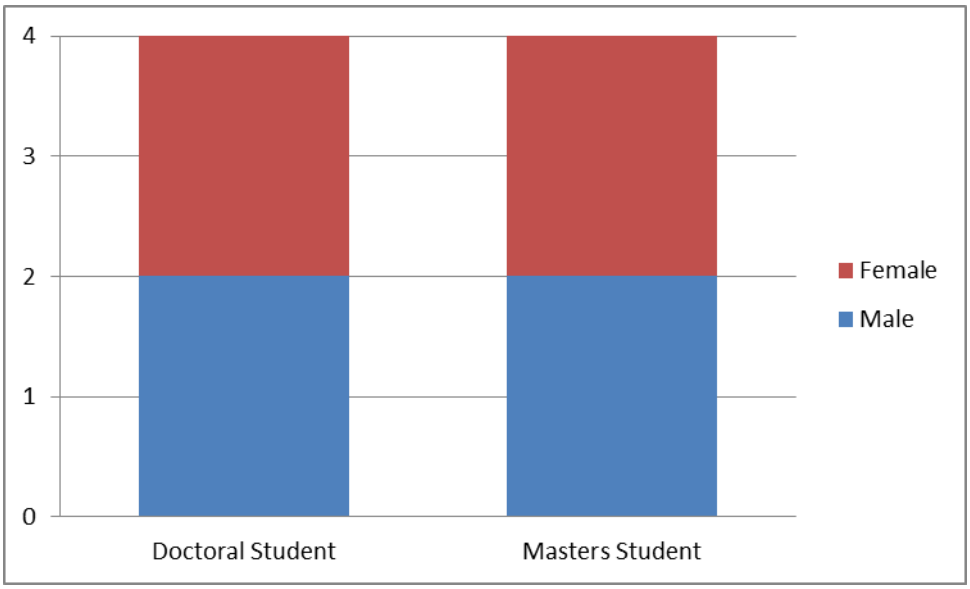

Figure1. Participants by Gender and Academic Level

They were in programs in four colleges (Arts and Sciences, Medicine, Architecture and Planning, and Engineering). From here on, I will group the participants from Engineering

Table1

Demographic Descriptors of Participants

	Academic Level		College Affiliation			Region of Origin		
Gender	Masters	Doctoral	Arts & Sciences	Medicine	Engineering/ Architecture & Planning	Americas	East Asia	South Asia
Male	2	2	2	1	1	1	1	2
Female	2	2	1	2	1	2	2	0
Total	4	4	3	3	2	3	3	2

and Architecture and Planning for all further descriptions or discussions as the number of international students in Architecture and Planning is so small as to be identifying (see Figure 2 for college affiliation).

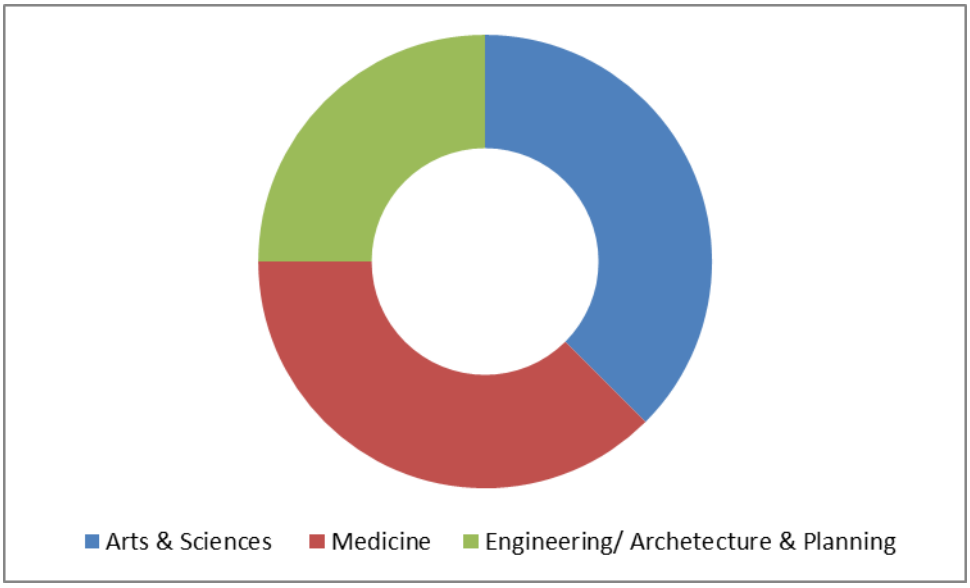

Figure 2. Participants by College of Graduate Program

The eight participants came from three geographic areas; three from the Americas, three from East Asia, and two participants from South Asia (see Figure 3 for region of origin).

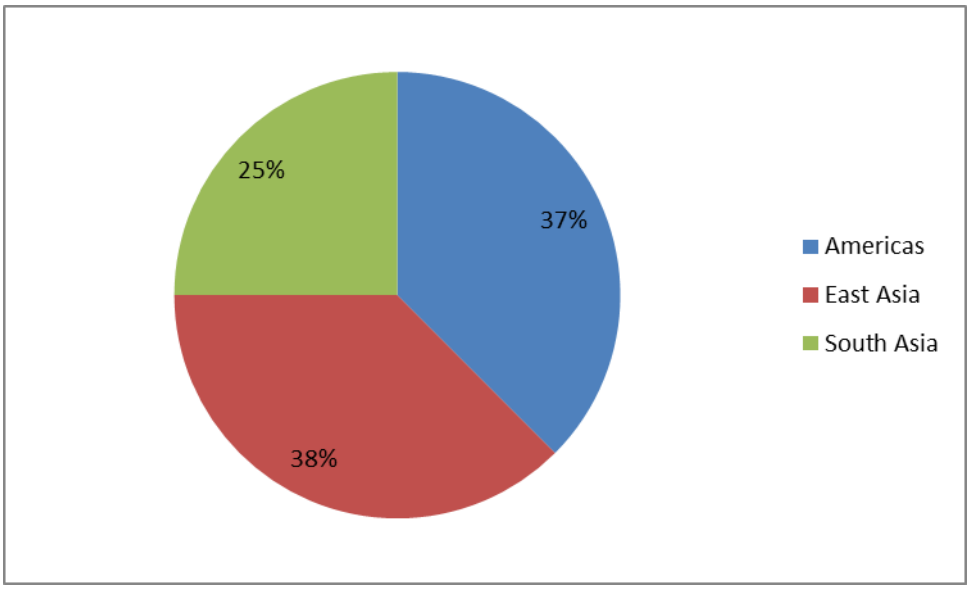

Figure 3. Participants by Region of Origin

Three of the participants had attended English medium of instruction elementary or secondary schools for all or part of their pre-college education. The participants do not represent the overall international graduate student population at this university, or the international graduate students at any college within this university; they represent themselves. They generously shared their time and stories through interviews that were conducted in a variety of locations of the participant's choice (coffee shops, study rooms in a university library, and a student center). Interviews lasted between 45 minutes and one hour and 45 minutes in duration.

My first steps in this analysis were to code nearly all meaning units in all eight interviews. I then combined those codes into progressively larger groupings ending when I had themes that I thought represented the meanings of the participants' stories as shared in their interviews. Following that process, I identified 212 total excerpts in the interview transcripts from my eight participants. I used 55 total codes (see Figure 4 for unique codes by theme),

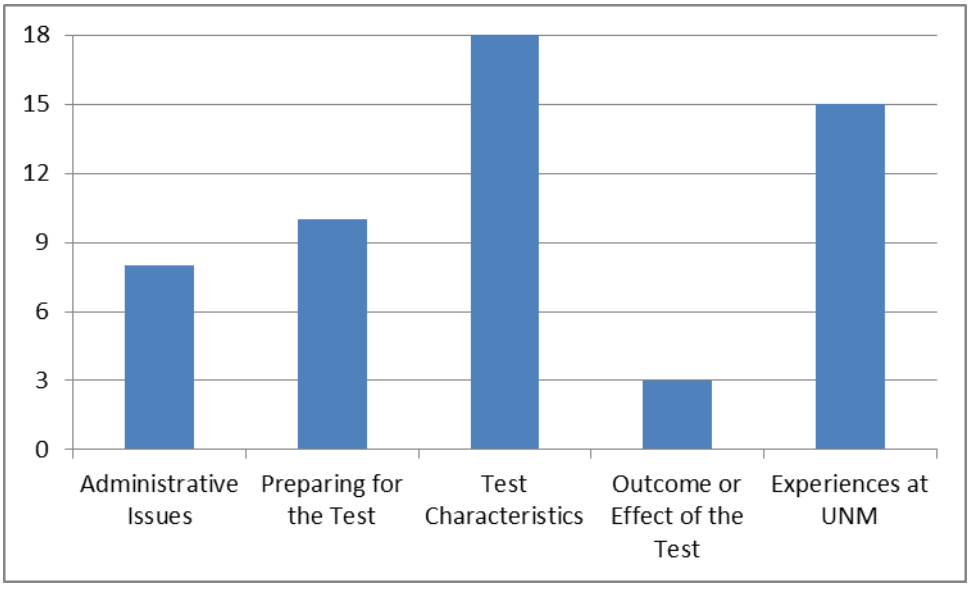

Figure 4. The Number of Unique Codes that Comprise Each Theme

and applied them a total of 624 times (see Figure 5 for code applications by theme).

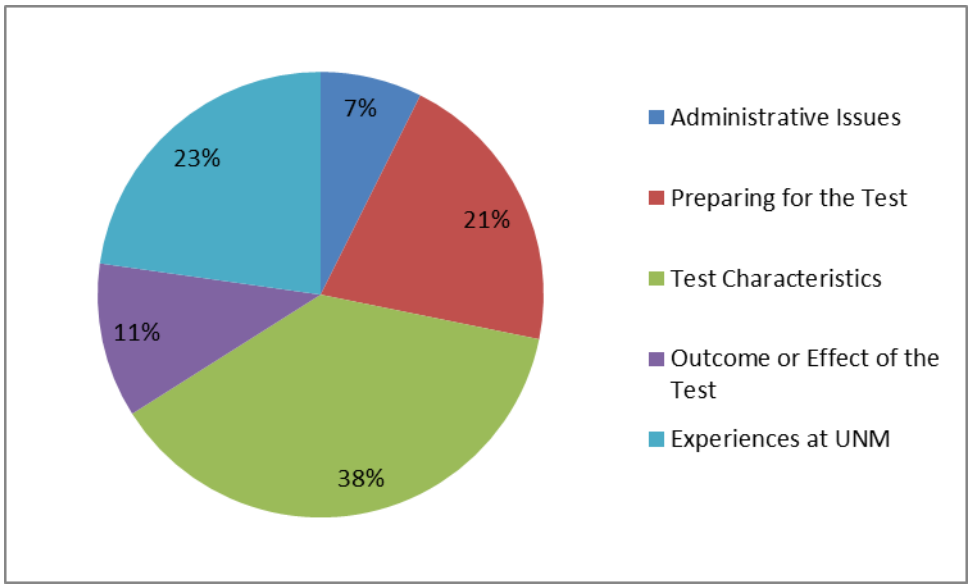

Figure 5. Percent of Total Code Applications by Theme

I organized the 55 codes into five themes wherein these five themes were composed of codes arranged in a hierarchical fashion, up to three levels deep (sub-themes, or first level codes; second level codes that were clustered under sub-themes, and third level codes that clustered under second level codes).

The themes that I identified were *Administrative Issues*, *Preparing for the Test*, *Test Characteristics*, *Outcome or Effect of the Test*, and *Experiences at UNM*. Four themes (*Administrative Issues*, *Preparing for the Test*, *Test Characteristics*, and *Outcome or Effect of the Test*) related to my first research question (What are the perceived experiences of non-native English speaking international graduate students with the internet-based version of the TOEFL?). Two themes, *Outcome or Effect of the Test*, and *Experiences at UNM* related to the second question (What are the students' perceptions of the applicability of the INB TOEFL in light of their subsequent experiences with academic English?). This combination of codes into the five themes was not the only

possible way to truly represent the participants' voices in the interviews, but it was the way that I chose to represent it for this analysis. All codes identified in this analysis were included in the themes, but not all codes were as meaning rich. Descriptions of the themes follow, with additional detail on those codes that most clearly revealed the participants' experiences and perceptions. See Table 2 for a list of all codes used in this research.

Administrative Issues

The *Administrative Issues* theme included eight codes at three levels. This theme included sub-themes (or first level codes) of 'ETS issues', and 'testing center issues'. 'ETS issues' had no second level codes. 'Testing center issues' was built from two second codes, and three third level codes. The second level codes were 'test center availability', and 'difficulties at the testing center', the later having three third level codes ('distractions at the testing center', 'general test center problems', and 'noisy'). The codes that clustered into the theme *Administrative Issues* were applied 46 times in this analysis. The codes that comprised the sub-theme 'test center issues' were applied a total of 39 times.

Administrative Issues related to the manner that ETS interacts with test takers in registration and communication, and the experiences participants had at the testing centers. Some participants had problems related to registration and communication with ETS; all participants reported some difficulties involving the testing center. The difficulties described included getting a seat at a nearby center, being distracted (often by noisiness at the center), or having other problems at the center.

Difficulties experienced by participants began with finding a seat at a local testing

Table 2

Codes used in this Research by Theme and Code Level

Administrative Issues	*Preparing for the Test*	*Test Characteristics*	*Outcome/Effect of the Test*	*Experiences at UNM*
ETS Issues	Nervousness	Test Process	Access to Higher Education	English Usage
Testing Center issues	Practice Exams	Types of Language Used in Test	Learning through the Test	*speaking most important
*Test center availability	Preparation for sections of the test	**everyday language	Comfort with English	*communication and interaction
*Difficulties at the testing center	*Reading	**formal language		*social
**Distractions at the testing center	*Writing	Differences and Difficulties		*Slang/ informal language
**General testing center problems	*Speaking and listening	**difficulty		*academic setting
**Noisy	**speaking	**impersonal lack of natural-ness		Writing in English
	**listening	**differences in testing process		*Reading in English
	**note taking	**differences in testing format		Experience in Initial Grad Classes
	**video/tv/radio	**domain of questions		*Faculty flexibility
		**speaking most important		*Differences in coursework expectations
		**computer		*Amount of reading
		Expectations of the Test		*required to interact or present
		*test is good		TOEFL Requirement
		*test bias		
		*sections/types of questions		
		*expected difficulty		
		*test score as expected		

*= 2nd level codes
**= 3rd level codes

center, covering the cost of taking the test, communicating with ETS, and difficulties or distractions once at the testing center. While few participants mentioned the cost of the test, it was mentioned as a difficulty by two participants. One participant was particularly effected by the expense of taking the test as he needed to take the test more than once. He also experienced some frustration in attempts at communication with ETS. Difficulties involving getting a seat at a local center or problems at the center effected more participants.

Difficulty getting a space at a local testing center was experienced by three of the participants. One participant shared that the test date and location that he had wanted filled up before he could get registered and that he checked the ETS web site every day hoping to find a seat was available. He shared that if he had not been able to register for the test at that location, he would have had to travel to another city several hours away in order to take the test. He was lucky, and was able to finally register for a seat at the local center. Another participant was not so lucky. She described this situation when she said:

> I had to travel to a different city to take the test. I live in a pretty big city in {home country}, but I had to go to another state. … I had a deadline for applying here so I didn't really have another option. I either took the test in a different city or I didn't get accepted into the program when I wanted to.

These participants were each living in large cities containing major universities, and in countries with large numbers of test takers. They began attempting to register for the test well in advance of the test administration date. Once registered, other issues could come up that might cause a delay in testing. One participant was able to register for her preferred testing center and date, but found that she was not allowed to take the test on

that date due to administrative rules. She shared that "I was going to take it one weekend and for some reason my name was not written correctly, so I had to come back the next weekend. So I was… I just wanted to get that over with." While less than half of the participants described experiencing administrative problems with registration or check-in, those who did experience these problems expressed a great deal of frustration with what should have been a simple process.

Once participants were allowed into the testing centers they found more problems. Most described difficulties and distractions at the testing centers. Describing a bad experience that a friend had, one participant said:

> I think some of the centers they … I think it's better off if they choose the centers well, because I know some of my friends they took the TOEFL at centers that were absolutely awful. They don't have proper computers, they don't have good access to do the exam. I think that is one big thing that I would want them to check, check more.

Several of the participants indicated that the headphones provided at the centers were not adequate given the closeness of the other test takers and the level of noise in the room.

The most frequent type of problem was related to distractions and noisiness. The participants did not indicate that the reading and writing sections of the INB TOEFL were adversely affected by the noisiness. However, problems with distractions or noise during the speaking and listening portions were discussed by all but one participant. One participant described the testing center as chaotic and loud. Another worried that her speaking score was artificially lowered because she stopped speaking a couple of times

due to the distraction of hearing other test takers speaking. That participant summarized the situation in the following:

> One of the things that I did remember that was really bad was the section where you have to speak, so the speaking section. You could hear the people around you. So, you were talking and trying to talk about something and somebody is immediately next to you trying to talk as well, it distracts you.

She suggested that better headphones would have made the experience much better, and less stressful. While the above described distractions may be the consequence of several people taking the test at the same time in relatively crowded spaces, other distractions that participants experienced were more avoidable. One participant told of testing center staff who chose to move chairs in the room while the test takers were working on the speaking and listening sections.

Some difficulties or distractions at the testing center were described by all but one participant. Participants shared that these problems lead to increased feelings of nervousness and even to fears that their scores would be affected by their reactions to these distractions. None of the participants indicated that they had contacted the ETS to report these situations.

Preparing for the Test

The theme *'Preparing for the Test'* was composed of ten codes. The four sub-themes that I combined into *Preparing for the Test* were: 'nervousness', 'practice exams', 'preparation for sections of the test', and 'other'. The second level codes that I combined into the 'preparation for sections of the test' sub-theme were: 'reading', 'writing', and

‘speaking and listening’ (that I created as a combination of the third level codes of ‘speaking’, ‘listening’, ‘note taking’, and ‘vide/tv/radio’).

I applied the codes that comprised *Preparing for the Test* a total of 130 times. I used ‘Nervousness’ 28 times, making it the single most used of the codes that contributed to *Preparing for the Test*. ‘Practice exams’ was applied 19 times, and ‘other’ was applied 16 times. I used codes related to ‘preparation for specific sections of the test’ a total 67 times.

Nervousness. Given the importance of this test for the test takers it was not surprising that all of the participants talked about their level of nervousness about the test. Some mentioned that they “weren’t too nervous” while others described nervousness almost to the point of fear. Nervousness was described during preparation for the test, while taking the test, and in the period between taking the test and receiving their score on the test. Because many of the stories that described nervousness took place during test preparation I have included it in this theme.

Nervousness related to high stakes testing is at least to some extent expected. A test taker’s score on the INB TOEFL can open doors, or shut them. One participant who told of generally feeling some nervousness or anxiety related to tests said that even compared to his usual “I wasn’t as comfortable as I normally am. Umm…so, so, that exam stress was definitely there.” He shared that this increased nervousness was related to the test being a “necessity” for admission to graduate school in the United States. Others who said they did not usually feel much or any anxiety around testing shared that they did with the INB TOEFL. One participant said that “before the examination I felt

very, very bad and nervous." Another participant related that the nervousness that he felt decreased once he began the test:

> So it is like initially it makes you a bit nervous. But once you are before the screen and you get the first question and you get the feel of oh this is something I can easily manage then you just start fighting back. It is just the first 5-10 minutes that you are … that give you a hard time. Later on once the exam is done you are good.

The possible effect of nervousness upon a test taker's score on the INB TOEFL was mentioned by some participants. One participant shared that she felt her score on the speaking section of the test may have been lower than it should have been due to her feeling nervous about the way that portion of the test was administered. Test takers' feelings of being in a physically and psychologically compromised state was something that one participant thought was important for the ETS to consider. She said that "people do need to consider the stress and the nervous as a very important part" of the experience of taking the INB TOEFL. She asserted her belief that the INB TOEFL not only assessed a test taker's skills in English but also "tests your ability to face the stress and work under high pressure." The ability to process academic English under stressful and distracting circumstances is not included in the description of the INB TOEFL, the implications of which will be discussed in Chapter Five.

Practice exams. All of the participants took some sort of practice exam. Most took advantage of the free practice exams offered online by ETS. This exam was valued by the participants and was the experience with ETS that seemed to be the most positive for the participants. Taking the practice exams not only gave the participants an

indication of how they would likely do on the exam, but also let test takers know “exactly what to expect.” One participant shared that he thought the experience of taking the ETS-provided practice exams would “help you to do the final TOEFL.” Another participant shared that she thought “the mock exams are really important” because they allow the test takers to be better “prepared for it and familiar with the situation.” All of the participants, including the three who had attended English medium of instruction schools during some part of their primary or secondary education, reported that taking the practice exams was useful to them.

Preparing for sections of the test. Participants mentioned practicing for all four sections of the exam, but most of the comments were about preparing for the speaking and listening sections of the assessment. Perhaps this is because, as some participants mentioned, the other sections (reading and writing) are more like other exams they have taken either in college, or for graduate school admission (i.e., the GRE). As one participant said “from the perspective of reading and understanding in English I was confident. I had a thorough experience with that. That was good. That was easier for me. As far as speaking goes… it was … not as easy.” These sections were described as more difficult to prepare for, not per se more difficult. Additionally, several participants shared that there were few if any native English speakers for them to interact with during test preparation, as one participant said “the environment isn’t so good because I didn’t have foreigners to speak to”, and that made preparing for and feeling confident about the speaking and listening sections less likely than feeling confident about the reading and writing sections.

Some participants found conversation partners to work with while they prepared for the speaking and listening sections of the test, but most did not. One participants said that because she did not have access to native English speakers she was pleased to be able to practice with someone whose spoken English was only a bit better than hers. Participants not only worked with other second language learners while preparing for these sections, the interactions also shaped their assumptions and expectations of the test. One participant said of his conversation partner that "I had a very good friend he is very excellent student, and his English is very good, and I practiced my oral English to him. He said you must express your idea clear, this is the first part, not the pronunciation." The extent that participants relied on other test takers for knowledge of what was expected varied, but half mentioned input from other test takers as part of their understanding of the test.

There seemed to be less surety about how the speaking section was scored than about how other sections were scored, and this made preparing for it more complex. One participant expressed this when she shared that:

> I know I can speak, and I guess if that is what they grade on you speak correctly, but academic level of speech was not very well. I could talk to people, and express myself, but not academic level. I could not easily use big words or have more formal speech.

This participant expected to easily pass the test, based both on her previous experience with education in English, and also on input from friends who had taken and passed the INB TOEFL. Because these friends knew her English language skill levels they were able to share with her that the INB TOEFL would not be hard for her. She had also done the

practice tests that ETS makes available online. Even with all of that, she was still somewhat unsure what was expected for the speaking section of the test.

Many of the participants said that they put most of their effort into preparing for the speaking and listening sections. One participant shared that she spent two hours every day for a few months preparing for the listening section of the test. Perhaps due to the minimal access many of the participants had to native or fluent English speakers, several of the participants described using television, movies, music and other media as part of their preparation for these sections of the test. Some shared that they learned more about English usage from the alternate sources than from classes, and also that they developed more connection to and interest in English through the alternate sources than through classes. One participant described her experience as:

> I didn't have a chance to really talk to native speakers. I tried my best to immerse myself in the English word. I put my ipod and listened to English music or VOA [Voice of America], BBC every day. If I really want to relax I just watch American movies and try to follow what they say.

Compared to studying for more traditional exams, she described this practice as a more relaxed experience as she did not need to "sit in front of a table and just read and recite". She used this input not just in preparation for the listening section of the exam, but also for the speaking section as she would "imitate what they say on TV what they sing in the lyrics" of the songs she was listening to.

The participants described their experiences preparing for the test as having both positive and negative aspects. The practice exams provided by the ETS were praised by all participants who used them. Participants put different amounts of time and effort into

preparing for the exam based on their perception of their English language skills, the requirements of the test, and the amount of time they had to prepare for it. All participants mentioned at least some intentional preparation for the exam.

Test Characteristics

Test Characteristics was comprised of 18 codes, or nearly one third of all codes used in this research, which was more than any other theme. *Test Characteristics* had four sub-themes (or first level codes) that were: 'test process', 'type of language in test', 'differences/difficulties', and 'expectations of test'. There were also second level codes for three of the four sub-themes. 'Types of language' was comprised of 'everyday language' and 'formal language'. 'Differences and difficulties' included 'difficulty', 'impersonal/lack of natural-ness', 'differences in testing process', 'differences in testing format', 'domain of questions', 'speaking most different/difficult', and 'computer'. 'Expectations of the test' was made up of 'test bias', 'test is good', 'sections/types of questions', 'expected difficulty', and 'test score as expected'.

The codes that comprised *Test Characteristics* were applied 236 times in this analysis. This was more code applications than for any other theme. I applied the code, or codes that comprised the sub-theme 'test process' 29 times, 'types of language in test' 44 times, 'differences/difficulties' 101 times, and 'expectations of test' 62 times.

Test process. The INB TOEFL was unlike other tests that the participants had taken prior to preparing to apply to graduate school. The participants described the reading and writing sections as similar to other tests, but the speaking and listening sections were, for them, unique to this test. The only other test like the INB TOEFL that the participants mentioned was the GRE. One participant talked about her experiences

with the IELTS, describing how different it was from the INB TOEFL. While some participants talked about the number of sections and the time allowed for each, those statements felt like an information transfer rather than a personal story. The more personal stories involved the unique aspects of the process of taking this test, from having to show an ID and having a picture taken, to not being allowed a bathroom break at one location, to being turned away because the name on the ID did not exactly match the name on the official list. The formality of the process was new to participants. One described the process as "you have to go to a special office where it is given, show your ID and everything is very formal. You can't have your things with you, not your phone or anything like that." She further related that bathroom breaks were not permitted except between sections of the test. One participant had taken the IELTS and described the rules for that test as very different, and more like a regular class test.

Types of language. The description of the type of language used in the INB TOEFL is one area where the participants varied. Most participants described the type of language used in the INB TOEFL as everyday language, rather than academic language. In contrast, some said that they felt it was more like academic English, meaning it was like the English used at the university. However, even those who said it was academic English allowed that the language used in the GRE was more like the language used in graduate school. Some participants shared that they felt that facility with both the types of English assessed by the INB TOEFL and the type of English assessed by the GRE were important for them as students at a U.S. university.

The participants who said that the INB TOEFL assessed everyday language not academic language described it as "day-to-day language", "language for general use",

language for “going to the store”, or “for general living here”. They relayed that their assessment of it as everyday language was due to the linguistic structure of the language used, the level of difficulty of the vocabulary used, and the types of questions asked on the INB TOEFL. One participant recalled that “there were examples of like a prof asking if you were going to be late for a meeting” and other similar examples that seemed to him to be everyday conversational English.

However, one aspect of everyday language that was reported as missing on the INB TOEFL was slang. One participant shared that he had trouble understanding his classmates during his first year as they used a lot of language that he described as slang. He said that slang of this type was not used on the INB TOEFL. This participant described the language used on the INB TOEFL as academic in part due to the lack of slang.

The participants all made comparisons between the INB TOEFL and the GRE such as; “I think it [the INB TOEFL] is more like ordinary conversation between people. I think TOEFL is not graduate level. GRE is graduate level. TOEFL is daily conversation level.” All participants, even those who described the English assessed by the INB TOEFL as academic English, described the language assessed by the GRE as more like the language used in graduate school. Some of the participants had clearly put a lot of thought into the two tests, their strengths, and differences. They offered detailed comparisons of the language assessed by the two tests such as

> GRE assesses you a lot on high end vocabulary compared to the TOEFL. TOEFL still has much simpler words. The words are more commonly used in day-to-day life. So as far as the academic setting goes, I guess GRE prepares you more from

> that perspective than TOEFL. TOEFL makes you express yourself quickly and easily, but when it comes to a much more commanding way of expressing GRE is second step up. It is more the academic. TOEFL is like the foundation. It is more basic. GRE is the next level.

The ability to read and write at a graduate level is important to the participants in this study. The ability to interact with spoken English is also important. The English assessed on these two tests were described as differing on structural complexity, difficulty of vocabulary, and subject of questions.

Differences and difficulties. The participants' comments spoke to the differences and difficulties of the test more than to any other sub-theme. Taken as a whole, the test was different from any other test they had taken; the types of questions, range of skills assessed, technology used in the assessment, and the impersonal nature of the test. Some of the differences lead to difficulties for the test takers. However, the test was generally described as not being very difficult, and as less difficult than other English language assessments that the participants had taken. While there was general agreement that the INB TOEFL was not a very difficult test, some specific sections (most often the speaking and listening parts) were identified as more difficult for some of the participants. The participants described these two sections as difficult due to problems at the testing center and the un-naturalness of these test sections more often than due to the content or complexity of the section.

None of the participants had ever taken a test like the INB TOEFL prior to applying for graduate school. Some took the GRE at about the same time as they took the INB TOEFL, and those who did shared that it was the only other test that they had taken

that was anything like the INB TOEFL. One participant shared that “it was something different to other exams definitely; the way you prepare for the exam, and the different sections that you have. It was uncommon to most of the exams that I take.” Some of the differences were related to the technology used in the administration of the test. The use of computer-based administration was unlike anything the participants had experienced previously. The impersonal nature of this administrative modality and the modality itself were identified as different and problematic. One participant stated that while computer-based tests were unknown to students of her generation, perhaps in the future students would be more accustomed to this type of assessment. Some of the concerns with the mode of assessment were related to uncertainty with the machine and accessories, and the lack of human interaction. One participant related:

> So it was compared to that [the IELTS], the experience was very different. It was impersonal, and you didn’t know how you did. It is a computer that you are talking to so it is different. Sometimes you don’t know if the mike is recording well what you are saying or not. You have to test the head phones. So in that sense you have more doubts than if your take a test direct, face to face with somebody. You can know how well or bad you are doing if you are talking face to face. The expression, you know tells you a lot. Talking to the computer, you don’t know.

Some participants shared that despite knowing what to expect with regard to sections and question types that they found the test to be challenging to take due to the test design. A participant shared his thoughts on this as:

> I mean as far as TOEFL goes you cannot just go and sit the test. You need to prepare yourself, and be ready for it, and have a mindset, so you are ready for any impromptu questions, any such situations. Definitely you have to orient your mind to the exam so as to be ready. You cannot just go randomly and say … that's what I think.

Impersonal/lack of naturalness. The impersonal nature of the assessment was mentioned by several of the participants. Several participants expressed frustration with this un-naturalness. One shared that he felt this was a big problem with the test, and that people need to be able to interact with other people, not computers. Others were concerned that their reaction to the unnaturalness of the process may have negatively affected their score on the test, most often their score on the speaking section. One participant said:

> I think that speaking to a person, or yeah, having a face to face interaction like Skype even, or that, would give a more accurate idea of how well a student is in English because talking to the computer makes you like very unnatural. I can see people getting stressed about that and nervous. And uh… others might just memorize something or else… though there is not really an interaction. In real talk there is. … You feel like you are just talking without response. And uh… you don't like there's no natural pause and there is no person to know whatever you are saying. You can stop talking for a little bit, but then you can lose track. With a person, you don't talk all the time like that.

Expectations of the test. Participants' expectations of the test were based on their experience with practice tests, previous English language exams, and what they were told

by friends and family members who had taken the INB TOEFL or previous versions of the TOEFL. They had expectations of the sections and types of questions, the difficulty of the test, and the score they thought they would receive on the test. Fortunately for the participants, their experiences were mostly in line with their expectations. They all said that the ETS-provided practice tests were very similar to the actual test. Therefore, they knew what the sections of the test would be, and the level of difficulty to expect. Most participants who talked about their expectations for their test score shared that the score they received was in the range that they had expected.

Sections/types of questions. Though the participants found that the practice tests informed them what to expect regarding the sections, timing, and types of questions, there were a few occasions when a participant was surprised by the test. Participants focused on different aspects of the test. Some were mainly concerned with the overall structure of the test, their comfort level with the material in the practice test, and their test score on the practice test. Others wanted more details about the test. One participant who wanted mainly general details reported that he recalled "knowing exactly what the sections were going to be… once you do the practice exam you know exactly what to expect, there was no surprise." He said that after receiving a passing score on the practice exam, and finding the material to be fairly easy, he was not too concerned about taking the actual test. Other participants focused in more on specific details of the test, such as one participant who "studied the structure, how many minutes there were for the parts, how many questions there would be, and which order they were going to come." After taking the practice exam and acquiring the level of detail that she felt comfortable with, she also was confident that she would get a good score on the test. Another participant

took one practice exam and also talked to a friend who assured her that the test would be easy for her, that she relayed was an accurate assessment.

Although all participants relayed that the practice tests informed them of the structure, order, timing, difficulty, and types of questions on the test, there were a few questions that were unexpected. A few participants reported being surprised by the subject of some of the questions. Another participant shared that she was not prepared for some of the topics that she was required to respond to in the speaking section of the test. She recalled that:

> the topic was very different. It was: describe a letter or an essay that you wrote to your mom. So, I wasn't prepared for that question. … I expected general - like what is your favorite sport, or favorite class, or favorite teacher. I was prepared for those questions, but not that.

The personal nature of the topic was part of what surprised her. She was also surprised by the specificity and subject of the topic. Another participant reported being surprised by a question about a U.S. holiday. She shared that she felt this was not appropriate.

Many of the other participants commented on the topics of the questions they received. Some participants shared happy coincidences such as getting a reading section on a topic that they were very informed (i.e., an area of study or of particular interest). Several participants shared general concerns about the appropriateness of some question topics given the range of fields of study of the test takers. Some described difficulty based on the subject of the questions. One participant had unexpected difficulty with one of the reading passages saying that "it took some time for me to understand the topic." He said that he had not had trouble with reading passages on the practice test, and felt that

the topic he had for one of his reading passages on the INB TOEFL was quite distant from any subject or topic that he had studied or had interest in.

Overall, participants found: the INB TOEFL to be what they expected based on the practice tests; their test score to be about what they expected; and the test to be at least "sufficient to be able to evaluate your English skills." They occasionally found questions that they had not expected or perceived as inappropriate. While the participants did not greatly praise the test, any praise from the test takers is likely an indication of success for the test designer.

Outcome or Effect of the Test

The theme *Outcome or Effect of the Test* had three sub-themes. The sub-themes were: 'access to higher education', 'learning through the test', and 'comfort with English'. The three codes were used 70 times. I used the code 'comfort with English' 28 times, 'access to higher education' 24 times, and 'learning through the test' 17 times.

Comfort with English. For some participants completing the INB TOEFL with a 'passing' score lead to them feeling more confident about their English language skills. Those who took the test multiple times or spent more time studying for the test were more likely to share that seeing a good score on the test had this effect. One participant said "yes, I felt my English was much better after I got this score. Perhaps this process is very slow. After I got this score I felt it, that my English was better." He further provided that he felt more confident talking to people, and felt that he would do alright at a U.S. school now after seeing a good score. Like others, he was particularly concerned with his oral English (speaking and listening), and so his score on those sections was most important in his increased confidence in his English language skills.

Learning through the test. Many participants also talked about how they had learned through the process of studying for and taking the test. One participant went so far as to suggest that people who want to learn English take the test even if they are not planning to come to U.S. schools because the process of the test itself was for her more useful than formal classes. In addition to suggesting that they had learned English through the process of the taking the test, some participants also allowed that the test provided some insight into how education would be at U.S. schools. One participant captured this idea as:

> like I said before, the TOEFL was very much related to the academic environment that you can expect. So, it is kind of helpful for an international student to take TOEFL to know what to expect in an American institution. So, it is kind of good preparation from that perspective. So when you actually come here you see, you feel it. It is something that you have been through some point in time during the exam or the preparation, so it prepares you in some way to face the situation in a more confident manner. So in that way it was helpful. It gave a big picture of what to expect. That is what I felt.

Not all participants believed that they learned through the process of preparing for and taking the test. Some made no comments related to this, and one explicitly said that he did not learn through taking the test.

Access to higher education. One effect of the test that some participants mentioned was access to higher education. Those who mentioned it expressed that it was very important to them. Some described it in a rather matter of fact manner, as one who said “TOEFL is a prerequisite. So at that point I was very keen to get into a good

academic institution so it was what I had to do, to take the TOEFL." One participant shared about this in more emotional terms. He said that while he had some issues with the test that it was very important for him and for other students like him, and that without it he would not be able to progress in his chosen field. He said:

> it provides a bridge for poor {country} students to come here. This chance is like 1 in 10,000 years chance. Now you can go from your hard work. Today's students want to … poor students want to make their family life better. Parents will get a high reputation in the community if they say 'My son is in US, studies there, got a PhD' they will get respect. It is also good for other students in your field. I am the second student to come here from my old university. So maybe the other students will … it is passing to next academic generation.

He shared that the TOEFL differentially offered this access to higher education. He said that in his country the IELTS is taken by wealthy students who were headed to schools that he did not believe offered scholarships to international students. He said that students in his country know that U.S. schools offer scholarships to international students, making it possible for students who come from poor families to study here. He described the INB TOEFL as a torment to be gotten through, but one that lead to greater opportunities and was thus worth the difficulties.

Experiences at UNM

The theme *Experiences at UNM* was related to my second research question. With 15 component codes, it was the theme with the second highest number of component codes. Its three sub-themes were 'TOEFL requirement', 'English usage', and 'experience in initial graduate classes'. 'English usage' included seven second level codes ('speaking

most important', 'communication and interaction', 'social', slang/informal language', 'academic setting', 'writing in English', and 'reading in English'). 'Experience in initial graduate classes' was comprised of four second level codes ('faculty flexibility', 'differences in coursework expectations', 'amount of reading', and 'requirement to interact or present').

I applied this theme's component codes 142 times. Two of the codes in this theme were each applied more than 20 times in this analysis. The sub-theme 'TOEFL requirement' was applied 29 times, and the second level code 'communication and interaction' was applied 24 times. The codes that comprised 'English usage' (including 'communication and interaction') were applied a total of 81 times. The codes that comprised 'Experience in initial grad classes' were applied a total of 32 times.

TOEFL requirement. Every one of the participants took the INB TOEFL. They took it because it was a requirement for admission into the graduate programs that they planned to enter. Therefore, it was not surprising that every participant shared something related to the 'TOEFL requirement'. Some participants directly commented on the requirement as when one shared that "the TOEFL is a prerequisite. So at that point I was very keen to get into a good academic institution so it was what I had to do, to take the TOEFL." Other participants shared less neutral stories related to the 'TOEFL requirement'.

Some participants expressed strong feelings about the requirement to take the TOEFL in order to be admitted to the university and to their chosen program. Those who expressed strong negative feelings about this requirement had some English medium of

instruction schooling prior to undertaking their undergraduate work. This frustration with the requirement was expressed by one participant when she said:

> I didn't feel like I really needed to take the TOEFL, although it was a requirement. I do understand why they make it a requirement, but knowing my background here, they should have been able to say she is going to be ok, she doesn't need to take the TOEFL.

Another suggested that the GRE should serve as adequate proof of English language proficiency, and that "if someone can pass the GRE then I am pretty sure he or she can pass the TOEFL too." This participant also reported feeling that "it is a waste of money" for applicants to have to take the INB TOEFL if they also have to take the GRE.

However, even those who expressed frustration with the requirement allowed that there was some utility in assessing a student's skills in communicating through spoken English. They recognized that the speaking and listening portions of the INB TOEFL are non-overlapping with the verbal section of the GRE. As a participant who reported feeling that while reading and writing were important, "the most important part is interaction between teacher and students, and the other students." This participant felt that getting a good score on the INB TOEFL showed prospective graduate programs that the student could interact with and through spoken English. A participant who voiced frustration about having to take the INB TOEFL said in relation to the INB TOEFL and the GRE:

> being able to communicate with classmates is more like the TOEFL. You need both of them. You are not going to talk to a friend informally in fancy wording, so you do need both of them. You need to know when to use which one, too.

This recognition of the importance of using English for communication and interaction with other people was referenced by all of the participants, and they shared that the INB TOEFL assessed this and the GRE did not.

English usage. Participants expressed that they used English directly, not through translation. They took notes in English, read class materials and literature for their research in English, and wrote their papers in English. Some even indicated that reading in their home language or writing papers first in their home language was more difficult than reading in English and writing in English. Participants also used English in most settings, both academic and social. Some participants had friends or family with whom they could converse in their home language. Most of them socialized with people who did not speak their home language on a frequent basis. One participant who found many conversation partners interested in speaking with her in her home language shared that she felt that she could express herself more authentically in her home language than in English even though she had used English in educations settings since elementary school.

One of the unique aspects of the INB TOEFL is the inclusion of speaking and listening sections. The skills that these sections assess were identified as important to the participants in their experiences as graduate students at this university both for interactions in formal academic settings, and in less formal settings. One participant shared that “it is very important for the professor to know that the oral and listening part are very good …so that you can communicate.” Assessment of oral communication skills were identified by the participants as strong points for the INB TOEFL. The participants reported needing facility with both academic and everyday English in their day-to-day life as graduate students.

In addition to talking about how important the ability to interact and communicate in English was to them, several of the participants shared stories of other international graduate students who had difficulty with using English in verbal communication. Some participants suggested that students from certain regions or language backgrounds had more trouble interacting in English. One participant mentioned that she knew "someone who got a really high score on the TOEFL, but when he came here he still has trouble to communicate with people." She suggested that when preparing for the INB TOEFL she was "more focused on passing the test" than on developing the skills needed to communicate with her classmates. However, she now considers that to be one of the more important skills that international graduate students need in order to be successful in their graduate programs.

Experiences in initial graduate classes. Most of the participants talked about how graduate courses here are different from their previous university work in their home countries. They described several differences from their previous educational experiences: differences in education model, differences in course loads and workloads per course, and differences in expectations of and interactions with faculty. Some of the participants expressed surprise at the differences, while others allowed that the INB TOEFL had implied these differences.

The participants described prior coursework as primarily lecture-based, where coursework in their current graduate program required more interaction including in class discussions, group work, and presentations. One participant shared that:

> the first year of schooling was tough because I had to take a lot of classes and be adjusted to the format because in my country it is different. We didn't have a lot of classes based on discussion and participation. It was mostly lectures.

This need to interact in class seemed to be a defining characteristic of course work here, and may have influenced their highlighting the need to communicate in spoken English (including both listening and speaking). One participant described his experiences as:

> In my country classrooms are based on lectures. Maybe now-a-days it changed a little. But in a lot of universities it is lecture based, so students do not have a lot of opportunities to participate in the class. So, here you require more participation, like presentation or discussion. So, it was a little difficult for me, also because of language because the classmates in my age use a lot of informal language… so sometimes when they talk about something I won't get it also because of cultural difference. I remember the first week was really difficult for me, was really tough. You had to really concentrate. …The reading amount was big for me.

Classroom interaction was not the only differences identified. Some participants also talked about other differences between their current program and their previous programs at their home universities. Some of the challenges described related to becoming settled in a new country, accustomed to different interaction styles within and outside of the classroom, and adjusted to what was often described as a greater course load and workload per course than they were previously accustomed to. One participant who had previously attended English medium of instruction schooling described the first semester here as:

> I definitely think that... having experience in college and school in {home country} are completely different. The perceptions once I got here … completely different. The amount of homework that you get here is extraordinary. In my country you do most of your work in the classroom. You have a few things to take home, but it doesn't even compare to the amount that you have to take home here. At home there are mostly lecture based classes. … We have more peer interaction [here].

Even with her previous experiences with the American education system this participant was surprised by the differences between her previous university experiences in her home country and the experiences and expectations in her current graduate program. For those participants who had not taken graduate level coursework at their home institutions some of the differences that they described may have been level-based; undergraduate classes being less interactive and requiring somewhat less intensive engagement with the course material than graduate classes. However, the participants who had prior graduate degrees also described similar situations, differences, and challenges.

Another aspect of their initial graduate classes that several of the participants described was their impressions of the expectations of their professors. Several allowed that the professors in their programs understood that new international students would not have polished English language skills. Some described programs that required less language intense courses in the first term. Others found their first term to be nearly overwhelming both in course domain and language requirements. Several shared that their professors were unexpectedly flexible, and one participant even hesitantly shared

that one of his professors helped to edit his early papers. This level of collegiality and support were mentioned as unexpected and greatly appreciated**.**

Concepts Crossing the Themes

The five themes arrayed in roughly chronological order from first contact with the ETS (registration for the test) through experiences in their graduate programs provide one glimpse into these international graduate students' experiences with the INB TOEFL and their subsequent English language use in graduate school. There are other concepts that intersect these themes. These concepts are connectivity, consequence, and communication. There were many commonalities across themes. Participants described a series of connected actions and events. These events involved communication among people and between people and agencies. These events had consequences, even life changing consequences for the participants. These concepts cross the themes, serving as the oft unseen warp threads upon which the decorative weft of the themes is woven.

In sharing the stories of their experiences with the INB TOEFL participants described connections between prior experiences and the assessment experience, among the sections of the assessment, and between their prior education experiences and current education experience. An example of this is that many participants spoke of the speaking and listening sections of the INB TOEFL almost as if they were one section. They also described connections between the other sections of the test. They shared about study practices that prepared them for multiple sections of the INB TOEFL.

The events described also had important consequences for the participants. One example is that problems with test registration could lead to long-term negative consequences for the participants if other alternatives were not available. If a seat could

not be secured at a local testing center a test taker risked missing an application deadline. One participant had to fly to another major city to take the INB TOEFL and meet the application deadline for graduate school. The connections and consequences were described by the participants in a ways that they would not and could not have been by other stakeholders.

A common idea running throughout the stories that the participants shared was the importance of communicating with others, or using English to communicate and interact with others. From problems communicating with the ETS, to an unwillingness to contact them when problems occurred at the testing centers, communication has direct implications for the test producers. Finding an appropriate communication partner, or finding media that spoke to the participant in some way was part of preparing for the test. Concern about if the rater could understand what was being said, and if the rater would understand the stress and distractions of the center was expressed by several participants. Being able to interact with their colleagues upon arrival at graduate school was identified as essential. Participants took the INB TOEFL because it was required, but developed skills in communication in English because they were essential.

The participants described connected events, not events in isolation, and also the need to connect with and communicate with other people throughout the process of preparing for the test, taking the test, and living into their chosen futures after the test. The participant's experiences have meaning and importance not only for themselves, but also for the test designers as they provide a different way of thinking about the assessment process. From this perspective one sees along a timeline that begins prior to registration and ends long after a 'passing' score is achieved.

Chapter 5

Discussion

The purpose of this research was to investigate the lived experience of test takers with the Internet-Based TOEFL. The specific questions addressed were:

1. What are the perceived experiences of non-native English speaking international graduate students with the internet-based version of the TOEFL?
2. What are these students' perceptions of the applicability of the internet-based TOEFL in light of their subsequent experiences with academic English?

These questions were addressed via a qualitative interview with eight international graduate students who had taken and passed the INB TOEFL.

This research shows that there is much to learn from the test takers. The stories shared by the test takers provide a window into the test taking process that is unique to them among all stake holder groups. Their experiences are important for those who would fully understand the test from the perspective of all stakeholders. All stakeholders add to the understanding of the test. However, the test takers are the only stakeholders who have the experience of preparing for the test, taking the test, and living with the consequences of the test. The stories shared by the test takers and the themes identified in those stories are meaningful to the test takers and they show the test in broader and more personally impactful context. The test is not an academic exercise; it is a gatekeeper, flinging wide the gate for those who achieve a 'passing' score, and slamming shut the gate for those who do not achieve a 'passing' score. The five themes identified in this research paint a picture of interconnected actions, hopes, and assumptions beginning long before the testing date and continuing long after the score was received. The events

connected to each other and consequentially to the participants' education and life experiences. Meaningfully, they also describe human connections, communication and interactions.

Summary of Findings

I identified five themes from the interviews. These themes were: *Administrative Issues, Preparing for the Test, Test Characteristics, Outcome or Effect of the Test,* and *Experiences at UNM*). The concepts of connections and consequence ran across these themes. *Administrative Issues* related to interaction with the ETS and the testing centers. *Preparing for the Test* related to studying for the INB TOEFL. This included prior experiences learning English and experiences with English such as listening to American music, or watching American movies and television programs either for entertainment or specifically in preparation for the exam. *Test Characteristics* included specific descriptors of the assessment comprised of sections and types of questions, difficulty level, expectations, and areas that were dissonant to the participants. *Outcome or Effect of the Test* told of the outcomes for the participants. They all passed and were admitted to a graduate program of their choice. Increased confidence with their English language abilities was one of the outcomes. *Experiences at UNM* was the final theme. It included the TOEFL requirement, English language usage at graduate school, and interactions with faculty and other colleagues.

In addition to the five themes I also saw three concepts that crossed themes. These concepts were connections, consequences and communication. The participants described a timeline that began long before sitting the test, and continued well after getting their test score. The connections between sections of the test, and particularly the

interconnectedness of the speaking and listening sections, were only part of this. Many participants described recreational activities as part of their preparation for the test. Participants did not describe discrete compartmentalization of academic or every-day language, receptive or expressive language, studying English or entertainment in English.

Consequences was another concept that crossed the themes. The purpose for taking the INB TOEFL was to get admitted to a graduate program at a US college or university. Participants described consequences related to every theme, and across time. Ultimately, all events and actions throughout the process of preparing for and taking the INB TOEFL all lead to the participants gaining entry into a graduate program of their choice.

Communication was the third concept that crossed themes. The participants described interaction and communication as primary needs throughout the process of preparing for and taking the INB TOEFL. Even those who felt that the requirement to take the INB TOEFL should be waive-able for those who have other documentation of English language skills shared that they felt the speaking and listening sections provided additional useful information for the admitting university.

Connections to Other Research

The test is described by ETS as developing within a "framework that takes into account models of communicative competence" (Cohen & Upton, 2006, p. ii). The participants in this research seemed to agree with the ETS that the INB TOEFL was a test of their ability to communicate in and through English. However they disagreed with the ETS as to the variant of English that the INB TOEFL assessed. Most of the participants found it to assess general purpose, or everyday language, rather than academic English.

There was some variation on this, with some participants describing the language used on the test as academic English, though all said that it was less descriptive of language used in graduate school than the language used on the GRE. Participants shared that both tests were important. The need to interact with others and use variants of English other than academic English was recognized by the participants. They recognized the importance of using everyday English for communication and interaction with other people both within and outside of the university. However, this mis-match between language variant assessed and language variant intended to be assessed is important. It speaks to the validity of the INB TOEFL as an assessment of academic English.

Structurally, the INB TOEFL contains four sub-sections. Factor analysis has suggested at least two different underlying structures (Stricker et al., 2009; Sawaki et al., 2009). Both of these analyses indicated that the speaking sub-test was least related to the other sub-tests, or to the higher order factor (in the one analysis in which a higher order factor was found). In contrast, the participants in this study often described the speaking and listening sections as highly inter-related.

The participants found the listening and speaking sections to be both the more important sections and also the most difficult to navigate. They reported difficulties with the format of the sections, including the equipment used, and the subject of the questions. Problems related to the subject of the questions were also reported for the reading and writing sections of the assessment.

Researchers have previously found interactions between test question subject and test takers' field of study (e.g., He & Shi, 2008; Kostin, 2004). These effects, while small (Kostin, 2004), could affect test scores (Sadighi & Zare, 2006). The participants shared

that questions on subjects distant from their fields of study were sometimes hard to follow, adding difficulty for them unrelated to the purpose of the test (assessing academic English). These texts failed to meet the standards set by Green, Unaldi and Weir (2010) as they are not comparable in domain or level, and therefore cannot be assumed to impose the same cognitive demands upon the test takers.

However, in this research even those participants who experienced frustrations based on the subject of the test questions allowed that the ETS could not produce field of study specific variants of the INB TOEFL. They expressed that getting questions well distant from, or occasionally related to, their field of study was just bad (or good) luck on their part. Even though any test score decrement due to test question subject is likely to be small, it brings into question the reliability of the test scores on the INB TOEFL.

More troubling were the stories of what I consider to be culturally inappropriate questions. Previous research has suggested that cultural and linguistic backgrounds are expected to influence test takers' spoken texts (Carey, Mannell, & Dunn, 2010; Chalhoub-Deville & Wigglesworth, 2005). The sort of cultural mis-match that the participants shared was beyond what I expected. Questions about interactions with specific family members were described by the participant who reported them as unexpected and inappropriate. In addition to cultural inappropriateness, there is also the possibility that those people may not be in the participant's lives (i.e., their mother may be deceased, and a question about last letter to her would therefore be unreasonable). Also, test takers may find questions related to discussions with family members to be too personal and not appropriate questions for an examination. Questions about U.S. holidays were described by some of the participants. I argue that U.S. holidays (i.e., Thanksgiving)

should never be considered acceptable subjects of test questions for the INB TOEFL as the target population for the INB TOEFL is non-U.S. students. Test designers and reviewers should be careful to avoid questions of these types as they impact on the fairness of the assessment.

Problems registering for a test administration were encountered by a quarter of the participants in this study. Distractions or other difficulties at the testing centers were reported by all of the participants. Some participants in this research mentioned that computer-based testing was unusual in their countries, and that the computer caused some difficulty for them. Some prior research has found that computer-based tests might cause difficulty for some test takers (Dooey, 2008; Maycok & Green, 2005). More often than the computers per se, participants in this research shared that the other equipment employed in performing the test (microphone, headphones, etc.) were problematic for them. Problems related to the specific equipment used while taking the test are unique to the test takers. No other stakeholder group experiences this aspect of the test.

The problems related to the other equipment went beyond simple use of the equipment. Participants shared that they were unsure if the microphones were working, and that they found speaking to a microphone rather than a person to be awkward and un-natural. This awkwardness was compounded by the awareness of others speaking at the same time. The headphones were described as insufficient for the environment; participants could hear other test takers speaking as well as general test center noises when wearing the headphones. Hearing the other test takers speaking and other general testing center noises, while they were trying to focus on the listening texts or produce their spoken texts were distracting for the participants. Some participants shared that they

were concerned that this combination of un-naturalness and concerns about the equipment may have led to lower than expected scores on the speaking and listening sections of the test. The method and mode of assessment was an intentional decision of the ETS in redesigning the TOEFL and creating the INB TOEFL. The un-naturalness of this can perhaps be mitigated through online practice exams. The choice of specific equipment used (microphone and head phones) is likely not based on careful design, but rather based on cost effectiveness. Changing to equipment more suited to the environment and test takers is in part a matter of knowing that there is a problem with the equipment and in part an economic decision. Noise cancelling headphones and microphones that provide visual feedback to the test takers that they are working might make the listening and speaking sections less problematic for the test takers. As "acceptance by test takers, test users, and the public is essential to the continued viability of the TOEFL" (Stricker & Attali 2001, p. 1) the cost associated with purchasing this higher quality equipment is likely to be well worth it to the ETS.

Additionally, nervousness above and beyond what participants described as normal for them affected all participants at some point during the process of preparing for, registering for, or taking the test. One participant asserted, the INB TOEFL "tests your ability to face the stress and work under high pressure" as much as it tests your knowledge of and facility with academic English.

Any influence of extraneous skills or abilities, such as the use of computers and management of high stress situations, offer challenges to the validity of the assessment. The effects of the computers may relax over time as test takers are likely becoming more comfortable with computers and therefore computer based assessment. Comfort with the

associated equipment, unless it also becomes more normal in other settings may not be expected to follow with increased comfort with computer and internet-based assessments. The ability to function under high stress situations may be seen in most environments, but should in my opinion be minimized whenever possible.

Despite describing some difficulties with the process or format of the test, participants said that the test was not very hard for them. Some specifically mentioned that it was easier than other tests of English that they had taken in college. He and Shi (2008) had previously found that their participants felt that the TOEFL was easier than other tests of English that they had taken. Most participants who mentioned the GRE specifically mentioned that the INB TOEFL was much easier than the English section of the GRE. As all of the participants in this research were graduate students this may have been an effect of their academic level. Undergraduate students may have a different perspective on the difficulty of the INB TOEFL.

All participants in this research were positively affected by taking the INB TOEFL as all were able to achieve a ‘passing’ score and gain entry into a graduate program in the U.S. in their fields of study. Additionally, some participants described learning through the test. Some said they learned more English, and even gained a greater feeling for the language in the process of preparing for the INB TOEFL. Others shared that they gained some perspectives, and expectations of how college would be in the U.S. by taking the INB TOEFL.

This research offers insight into the experiences of a group of graduate students who achieved a ‘passing’ score on the INB TOEFL. It offers no insight into the experiences of those test takers who did not achieve a ‘passing’ score on the INB

TOEFL. Their experiences with the test, opinions about the test, and the effects of these experiences upon their lives may likely be quite different. Previous research (Tsai & Tsou, 2009) found that test taker's opinions about an assessment varied. Those who had better language skills and who reported feeling less test-related stress had more positive opinions about the assessment.

Similarly, undergraduate students may have very different experiences with the INB TOEFL and those experiences may have very different effects upon their lives. In this research, comments about the difficulty and variant of English used in the INB TOEFL often referenced the GRE as a point of comparison. Undergraduate students may in particular have different opinions about the difficulty of and variant of language used in the INB TOEFL as they will not have the GRE as a point of comparison. This research offers no insight into the experiences of undergraduate student test takers.

Once in their graduate programs all participants described using English for all academic tasks, even those (like note-taking and writing first drafts of papers) that could have been carried out in their native languages. They shared that their professors were understanding of their process of acclimation to English medium of instruction classes. Often they said that their professors allowed some flexibility in the first term courses. However, they also described their professors as having high standards and high expectations for them. This included the expectation that they would progress in their written and spoken English as well as skills and knowledge in their field of study.

These experiences are unique to this stakeholder group. No other stakeholder group experiences the INB TOEFL in the same way as the test takers.

While researchers can hire subjects to take the test, they cannot create consequential conditions that mirror those experienced by these test takers. Their test takers' experiences, as described by the participants in this research, are of taking a high-stakes test under often less than optimal conditions. Frequent and detailed follow-up with test takers is the only way to learn of their experiences with the INB TOEFL.

The test, designed and produced by a US corporation, carries assumptions that are at odds with the assumptions of the participants. The participants had not taken another test like the INB TOEFL prior to application to graduate school. The test is a variant on the standard US psychometric assessment, employing technology not uncommon to US students. The participants, however, were often not accustomed to this sort of test and many of the participants were not comfortable with the testing modality. Some were faced with questions that they found to be inexplicable and inappropriate on a formal assessment. Others described unreasonable conditions at the testing centers. Interestingly, none of the participants contacted the ETS about these or any other issues (such as the distractions or difficulties at the testing centers). This research suggests that following-up with the test takers at different point in the process of preparing for and taking the test may yield information useful to the ETS including information about centers they use, and the difficulties that students experienced unrelated to their mastery of academic English.

The issues raised by the participants in this research suggest some challenges to the validity, reliability, and fairness of the INB TOEFL. Use of psychometric tests is underpinned by the assumption that they are reliable, valid, and fair. Therefore, these challenges must be taken seriously. Eliciting information from test takers about their

experiences is important if the ETS wishes to expand beyond the top down psychometric approaches that have characterized the design of large scale high-stakes test (Fairbairn & Fox, 2009). Further, information elicited from test takers provides a different view of the assessment than information elicited from other stakeholder groups. As the group with arguably the highest stake (Hamp-Lyons, 2000), test taker experiences and perspectives should, in my opinion, be taken seriously, and intentionally collected on a regular basis. Stickler and Attali (2010) called for more studies of INB TOEFL test takers and their attitudes towards the test. I believe that eliciting and understanding their experiences is essential to understanding their attitudes. This research suggests that direct methods of eliciting opinions and experiences will be needed, as even those participants with serious concerns about some aspect of the INB TOEFL reported not contacting the ETS about their concerns. Survey methods can give a wide but shallow picture of the test taking experience; interviews can give a deep but narrow picture of the experience. Both are needed to get a complete picture of this experience, and therefore understand the test takers' attitudes, needs, and concerns.

Limitations of the Study

Limitations of this study are primarily related to paradigm, participants, methodology, and researcher characteristics. As qualitative research this study does not present findings that are generalizable as quantitative research would, but rather presents findings that are potentially transferable. The stories shared by the participants in this study are their own and not claimed as anything other than that. The extent to which the results of this research is transferable to other settings or people will depend on how similar these settings and people are to those of this study. Consideration of the

characteristics of the environment and participants is therefore essential when considering the extent that these findings are transferable. It is important to state that while there are limitations related to participants, the participants themselves are not limitations of this study, but characteristics of them as a group may limit the transferability of the study to other groups and the interpretation of the findings suggested in this research.

All of the participants in this research received passing scores and were able to enter into graduate programs of their choice. Only one participant took the INB TOEFL more than once. The experiences of other groups (i.e., undergraduate students, anyone who never received a 'passing' score on the exam, test takers who were not able to enter a program at a US college or university) may be very different from the experiences of the participants in this study. All of the participants had taken the exam at least two years prior to sharing their stories. This temporal element may be meaningful; stories shared during the process or soon after completion of the exam may be qualitatively different from stories shared years after taking the INB TOEFL.

The number of participants was low, at only eight. These participants were not representative of the overall graduate student population at this university with regard to country or region of origin, academic program or college, or degree level. Other characteristics were not recorded, such as age, number of languages spoken, or INB TOEFL test score, and therefore nothing can be known about how they compare to the general population of graduate students at this university. An additional limitation related to participants is that only two opted to take part in the member check.

Limitations due to methodology are the data collected, analysis performed, iterations and change of the protocol, and lack of triangulation. Limited data was

collected on participant characteristics (as described above) and no correlations or other quantitative analysis was performed on the characteristics that were collected and themes found in this research. INB TOEFL test score was not collected therefore no correlation of themes and test performance could be conducted. No changes were made to the protocol across participants, and no additional data was collected for the purpose of triangulation.

Limitations related to the research are likely many. Some of these potential limitations include the researcher's minimal prior experience with qualitative interview research, lack of adequate planning for difficulties in finding participants, and failure to identify biases. While biases that I recognized were discussed with a critical friend, I could not have identified all of my biases.

Future Possibilities

This research suggests that the ETS (and by extension all test design and administration companies) should seek out the experiences and opinions of their test takers. The participants in this research all reported experiencing non-optimal conditions at the testing center and a variety of other problems with the test, but none of them reported bringing these problems to the attention of the ETS. These conditions likely affected other test takers as well as these participants. Based on these findings, the ETS might therefore consider implementing follow-up surveys and interviews with test takers. All participants in this research were successful with the INB TOEFL and in their further education. This research, illuminating problems even for this successful group of test takers, foreshadows the possibility of even greater problems for other groups of test

takers. Eliciting the experiences of those who scored well on the test and also those who did not could reveal commonalities and differences in experiences of these two groups.

Further afield, this research suggests that follow-up with test takers is beneficial in general. Test researchers, test designers and test administrators might interview people who have taken other large scale high-stakes standardized tests (such as American students who have taken the GRE). Eliciting the experiences of teachers on the use and administration of standardized tests in their classroom is also suggested.

The participants in this research had not previously experienced this sort of disconnected, impersonal assessment, but young American school children experience high-stakes standardized testing from the earliest years in public education as the American education has increasingly moved to a commercial product model from a public service/educational model. With such high stakes, failing to include all groups can lead to negative consequences for the test takers, administrators, teachers, and others. Including as many stakeholder groups as possible in discussion of the design and implementation of assessments could not only mitigate these negative consequences, but lead to better test design. Failing to include as many other groups as possible can lead to designers experiencing a sort of echo chamber, wherein all those included in the conversation have similar experiences, education, perspectives, and expectations.

I invite and encourage test designers to recognize the unique experiences and perspectives of the test takers and begin to include them more often in the design and analysis of assessments. This research suggests that there is much information that this stakeholder group can offer. It also suggests that they need to be asked to share their

experiences and opinions, as they are likely disinclined to contact the testing agency with complaints.

Appendices

Appendix A

Free Association Task

1. ‘Please share four words that describe your last many months since you took the internet-based TOEFL.’
2. I will repeat back these four words one at a time. After I say each one, please respond with any words or phrases that come to mind when thinking about that term.’
3. ‘When I say <first word > what words of phrases come to mind?’
    a. ‘Do you want to add any other words?’
4. ‘When I say <second word > what words of phrases come to mind?’
    a. ‘Do you want to add any other words?’
5. ‘When I say <third word > what words of phrases come to mind?’
    a. ‘Do you want to add any other words?
6. ‘When I say <fourth word > what words of phrases come to mind?’
    a. ‘Do you want to add any other words?’

‘Thank you.’

Appendix B

Interview Questions

Thinking about your experiences with the internet-based TOEFL, and academic English usage:

1. Please tell me about your experiences taking the internet-based TOEFL.

(Possible probes):

Describe your experiences taking the TOEFL.

What happened and how did it make you feel?

Tell me what it was like to take the TOEFL.

Can you tell me anything else?

What else happened?

Can you tell me about your response?

2. How do you use English in your academic work here?

(Possible probes):

In classes,

In labs,

In writing

In reading,

In communication with faculty,

In projects with other students,

In studying with other students,

In your studies?

3. Now that you have been at UNM for a while, do you have any thoughts about the TOEFL?

(possible probes)

Did anything about the test lead you to have assumptions about faculty or colleague expectations about your use of English?

Do you think the INB TOEFL give an accurate assessment of your skills in English?

Did the INB TOEFL meet your expectations?

4. To what extent do you think the TOEFL assessed the kind of English used in graduate school?

5. Is there anything else you would like to add?

(possible probes)

If the topic of test preparation was not mentioned yet, query with

How did you prepare for the test?

Is there anything you wish the test designers knew about the test?

Thank you for your time

Appendix C

Member check questions

I would like to share with you what I have found in the interviews I have done on the TOEFL assessment and academic English.

1. I found XXXXX theme.
2. I believe that these statements from your interview are related to this theme.
3. Do you agree/disagree that these statements are related to this theme?
4. Please explain.

(Repeat questions 1-4 for each theme.)

Is there anything else you would like to add?

Appendix D

Participant Recruitment Protocol

I am a doctoral candidate researching experiences of international graduate students who have taken the internet-based TOEFL. If you are a current student in a graduate program unrelated to language development or theory, have taken the internet-based TOEFL, are in your second or subsequent semester at UNM, please contact me if you are interested in participating in my dissertation research. Participation will take about one to one and a half hours, and will primarily include an interview. After my analysis is completed I will requests a brief (about 30 minute) follow up to confirm the accuracy of my analysis. No payment will be provided to participants.

If you are interested please contact me at 505.301.0708 or annalies@unm.edu

Appendix E

Transcriptions Conventions

Direct references by a participant of their home country, home region or state, or prior university was excluded from transcription when possible. Other information that was not transcribed included names of professors, advisors, and educational or funding agencies. Home or native language was also not transcribed. Examples of when these statements were not transcribed include statement of the type ‘I am from Albuquerque, New Mexico’, ‘I attended the University of Texas at El Paso when I took the GRE’, ‘other American students find this sort of test to be difficult’, or ‘for native French speakers pronouncing an initial H has to be learned’.

When reference were internal to other statements that I felt need to be transcribed (i.e., ‘I was living in El Paso Texas when I applied for the GRE, but could not get a seat for the test there, even though El Paso is a big city, so I had to go to Dallas to take it’, ‘in the U.S., we are accustomed to high-stakes standardized test, but I had never previously taken a test that made me feel so anxious’) I placed an indicator of the elided personal information within curly brackets. For instance I would have transcribed the previous example as ‘I was living in {large city} when I applied for the GRE, but could not get a seat for the test there, even though {city} is a big city, so I had to fly to {other large city} to take it’.

The University of New Mexico
Consent to Participate in Research

A Thematic analysis of experiences of non-native English speaking international graduate students with the internet-based Test of English as a Foreign Language

Introduction

You are being asked to participate in a research study that is being done by Annaliese Mayette, who is the Principal Investigator from the Department of Language, Literacy & Sociocultural Studies. This research is studying the experiences and perceptions of people who have taken the internet-based TOEFL and their subsequent experiences with English in an academic setting.

Previous research into test taker needs, perspectives and experiences has been quite limited. This study will expand on the limited research in the field. A few recent studies have addressed test-taker experiences and perceptions of the TOEFL, the results of which suggest that further research into test taker experiences is needed. Research into these experiences may inform test design, administration, and interpretation.

You are being asked to participate in this study because (a) You are a non-native English speaking international graduate student, (b) in your second or subsequent semester at the University of New Mexico, (c) you have successfully taken the internet based TOEFL , and (d) your program of studies, research focus, and previous experiences are in domains other than language development or language teaching. Eight to 12 people will take part in this study at the University of New Mexico.

This form will explain the research study, and will also explain the possible risks as well as the possible benefits to you. We encourage you to talk with your family and friends before you decide to take part in this research study. If you have any questions, please ask one of the study investigators.

What will happen if I decide to participate?

If you agree to participate, the following things will happen:

Your participation will include an interview and a free association task which together are expected to be of approximately one to one and a half hours in duration, as well as a second brief interview once my initial analysis is complete that is expected to be of approximately one half to one hour in duration. The purpose of the second interview is to share my findings with you and ask for your input on them.

How long will I be in this study?

Participation in this study will take a total of two to two and a half hours over a period of two days several weeks apart.

What are the risks or side effects of being in this study?

Risks of participation are no greater than risks associated with discussions of a personal nature with friends. You might become uncomfortable discussing personal, possibly unpleasant aspects of your experiences.

There are risks of stress, emotional distress, inconvenience and possible loss of privacy and confidentiality associated with participating in a research study.

For more information about risks and side effects, ask the investigator.

What are the benefits to being in this study?

Potential indirect benefits to you include being part of a group who helped to fill the gap in assessment design research, and potentially influence designers to consider the voice of test takers. Additionally, I hope that you find sharing your story to be empowering.

What other choices do I have if I do not want to be in this study?

You do not have to participate in this study. You may terminate participation at any time.

How will my information be kept confidential?

We will take measures to protect the security of all your personal information, but we cannot guarantee confidentiality of all study data.

Information contained in your study records is used by study staff. The University of New Mexico Institutional Review Board (IRB) that oversees human subject research and/or other entities may be permitted to access your records. There may be times when we are required by law to share your information. However, your name will not be used in any published reports about this study.

I will assign a pseudonym to you, and use that pseudonym in all communication about you and your responses including in my dissertation text. I will not transcribe any information that you share during the interview that is of a private nature and is not directly pertinent to my research. Further, I will disguise any relevant information that you share about specific location in country of origin (region or city) and prior institution of attendance.

What are the costs of taking part in this study?

There is no cost to take part in this study.

Will I be paid for taking part in this study?

You will not receive any compensation for participation in this study.

How will I know if you learn something new that may change my mind about participating?

You will be informed of any significant new findings that become available during the course of the study, such as changes in the risks or benefits resulting from participating in the research or new alternatives to participation that might change your mind about participating.

Can I stop being in the study once I begin?

Your participation in this study is completely voluntary. You have the right to choose not to participate or to withdraw your participation at any point in this study without affecting your future health care or other services to which you are entitled.

Whom can I call with questions or complaints about this study?

If you have any questions, concerns or complaints at any time about the research study, Annaliese Mayette will be glad to answer them at 505.301.0708. Dr. Julia Scherba de Valenzuela (Annaliese's faculty advisor) can also be contacted at 505.228.3450 Monday to Friday 10 am to 5 pm.
If you need to contact someone after business hours or on weekends, please call 505.301.0708 and ask for Annaliese Mayette.
If you would like to speak with someone other than the research team, you may call the UNMHSC HRPO at (505) 272-1129.

Whom can I call with questions about my rights as a research participant?

If you have questions regarding your rights as a research participant, you may call the UNMHSC HRPO at (505) 272-1129. The HRPO is a group of people from UNM and the community who provide independent oversight of safety and ethical issues related to research involving human participants. For more information, you may also access the IRB website at http://hsc.unm.edu/som/research/hrrc/irbhome.shtml.

CONSENT

You are making a decision whether to participate in this study. Your signature below indicates that you read the information provided. By signing this consent form, you are not waiving any of your legal rights as a research participant.

I have had an opportunity to ask questions and all questions have been answered to my satisfaction. By signing this consent form, I agree to participate in this study. A copy of this consent form will be provided to you.

_____ _____ _____

Name of Adult Subject (print) Signature of Adult Subject Date

INVESTIGATOR SIGNATURE

I have explained the research to the participant and answered all of his/her questions. I believe that he/she understands the information described in this consent form and freely consents to participate.

_____

Name of Investigator/ Research Team Member (type or print)

_____ _____

(Signature of Investigator/ Research Team Member) Date

References

Alderson, C. (2009). Test of English as a Foreign Language Internet-based Test (TOEFL iBT). *Language Testing, 26*(4), 621–631.

Atkinson, R., & Geiser, S. (2009). *Reflections on a century of college admissions tests* (Research Report No. CSHE.4.09). Research and Occasional Paper Series (p. 1-21). Berkley, CA: University of California, Berkley. Retrieved from http://cshe.berkeley.edu/publications/docs/ROPS-AtkinsonGeiser-Tests-04-15-09.pdf

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

Batmale, L. (1948). Veterans' high-school graduation by examination. *The School Review*, *56*(4), 229–235.

Beckett, L. (2011, September 19). America's most outrageous teacher cheating scandals. *Pro-Publica*. Online. Retrieved from http://www.propublica.org/article/americas-most-outrageous-teacher-cheating-scandals

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (Research Report No. MS - 25: RM-04-03). TOEFLS Monograph Series. Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-04-03.pdf

Bracey, G. (2005). The 15th Bracey report on the condition of education. *Phi Delta Kappan, 87*(2), 138–153.

Brown, G., & Hirschfeld, G. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy and Practice, 15*(1), 3–17. doi:10.1080/09695940701876003

Carey, M., Mannell, R., & Dunn, P. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219. doi:10.1177/0265532210393704

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383–391.

Chapelle, C., Grabe, W., & Burns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (Research Report No. RM-97-03: MS-10). Monograph Series. Princeton, New Jersey: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RM-97-03.pdf

Clarke, M., Madaus, G., Horn, C., & Ramos, M. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, (2), 159–181.

Cohen, A., & Upton, T. (2006). *Strategies in responding to the new TOEFL reading task* (Research Report No. MS - 33: RR-06-06). TOEFL Monograph Series. Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/research/policy_research_reports/rr-06-06_toefl-ms-33

Cohen, A., & Upton, T. (2007). "I want to go back to the text": Response strategies on the reading subtest of the new TOEFL. *Language Testing*, *24*(2), 209–250. doi:10.1177/0265532207076364

Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Crystal, D. (2000). Emerging Englishes. *English Teaching Professional*, *14*, 3–6.

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176.

DeMerle, C. (2006, December). *A multi-state political process analysis of the anti-testing movement* (Dissertation). University of North Texas.

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, *20*(1), 21–34. doi:10.1017/S0958344008000311

Educational Testing Service (2003). Test review: The TOEFL CBT (Computer-based test). *Language Testing*, *20*(1), 111–123.

Educational Testing Service. (2011). *Information for score users, teachers and learners* (Research Report No. Series 1, Volume 5). TOEFL iBT Research Insight. Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v5.pdf

Enright, M. (2004). Research issues in high-stakes communicative language testing: Reflections on TOEFL's new directions. *TESOL Quarterly*, *38*(1), 147–151.

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, *27*(3), 317–334. doi:10.1177/0265532210363144

Fairbairn, S., & Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: Essential considerations for test developers and decision makers. *Educational Measurement: Issues and Practices*, *28*(1), 10–24.

Fulcher, G. (2007, April 13). Universities undermine their own foundations: Contracting out English preparation courses in Britain is a short-term fix. *Guardian*. London. Retrieved from http://education.guardian.co.uk/tefl/comment/story/0,,2055735,00.html

Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy & N. Hornberger (Eds.) *Encyclopedia of language and education (pp. 157-176)*. NY, NY: Springer Science and Business Media LLC.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge Taylor and Frances.

Garrison, M. (2009). *A measure of failure: The political origins of standardized testing*. Albany, NY: State University of New York Press.

Gillan, M., Damachis, B., & McGuire, J. (2003). Australia in India: Commodification and internationalization of higher education. *Economic and Political Weekly, 38*(14), 1395–1403.

Goh, D. (2004). *Assessment accommodations for diverse learners*. Boston, MA: Allyn & Bacon.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*(3), 417–444. doi:10.1177/0265532207077209

Grace, G. (1989). Education: Commodity or public good? *British Journal of Educational Studies*, *37*(3), 207–211. doi:10.1080/00071005.1989.9973812

Green, A., Unaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*(2), 191–211. doi:10.1177/0265532209349471

Haertel, E., & Herman, J. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the national society for the study of education*, *104*(2), 1–34.

Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education*, *74*(5), 262–273.

Hale, G., Stansfield, C., & Duran, R. (1984). A comprehensive TOEFL bibliography 1963-1982. *The Modern Language Journal*, *68*(1), 45–51.

Halic, O., Greenberg, K., & Paulus, T. (2009). Language and academic identity: A study of the experiences of non-native English speaking international students. *International Education*, *38*(2), 73–93.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, *28*, 579–591.

Hamp-Lyons, L., & Lynch, B. K. (1998). Perspectives on validity: a historical analysis of language testing conference abstracts. In A. Kunnan (Ed.), *Validation in language assessment: selected papers from the 17th Language Testing Research Colloquium* (pp. 253–276). Mahwah, NJ: Lawrence Erlbaum.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assessing Writing*, *13*(2), 130–149.

Huang, J. (2004). Voices from Chinese students: Professors' use of English affects academic listening. *College Student Journal*, *38*(2), 212–223.

Huang, J. (2006). English language abilities for academic listening: How confident are Chinese students? *College Student Journal*, *40*(2), 218–226.

Hughes, A. (2003). *Testing for language teachers (2nd ed.)*. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Hursh, D. (2007). Assessing No Child Left Behind and the ride of neoliberal education policies. *American Educational Research Journal*, *44*(3), 493–518.

Hussar, W., & Bailey, T. (2008). *Projection of Education Statistics 2017* (Research Report No. NCES 2008078). Projections of Education Statistics. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2008/2008078.pdf

Institute of International Education. (2010). *Fall 2010 International Student Enrollments Survey* (Research Report). Open Doors International Student Enrollment Survey. Chicago, IL: Institute of Internationals Education. Retrieved from http://www.iie.org/en/Who-We-Are/News-and-Events/Press-Center/Press-Releases/2010/~/media/Files/Corporate/Publications/2010-Fall-2010-International-Student-Enrollment-Survey-Report.ashx

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. doi:10.1093/applin/amm017

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 Framework: A working paper* (Research Report No. RM-00-03, TOEFL-MS-16). TOEFL Monograph Series. Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-00-03.pdf

Jung, Y., Pawlowski, S., & Wiley-Patton, S. (2009). Conducting social cognition research in IS: A methodology for eliciting and analyzing social representations. *Communications of the Association for Information Systems, 24*(Article 35). Retrieved from http://aisel.aisnet.org/cais/vol24/iss1/35

Kachru, B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literature* (pp. 11–36). Cambridge: Cambridge University Press.

Kachru, B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching, 25,* 1–4.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koll, O., von Wallpach, S., & Kreuzer, M. (2010). Multi-method research on consumer-brand associations: Comparing free associations, storytelling, and collages. *Psychology and Marketing*, *27*(6), 584–602. doi:10.1002/mar20346

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TEOFL dialogue items* (Research Report No. RR-04-11). Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/research/policy_research_reports/rr-04-11

Koyama, J. P. (2011). Generating, comparing, manipulating, categorizing: reporting, and sometimes fabricating data to comply with No Child Left Behind mandates.

Journal of Education Policy, *26*(5), 701–720. doi:10.1080/02680939.2011.587542

Krashen, S. (2011, June 5). *Our schools are not broken: The problem is poverty*. Commencement address presented at the 2011 College commencement, Graduate School of Education and Counseling, Lewis and Clark College. Retrieved from http://www.substancenews.net/articles.php?page=2319&section=Article

Kunnan, A. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge: Cambridge University Press.

Kunnan, A. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*(27), 183–189. doi:10.1177/0265532209349468

Lahlou, S. (1996). A method to extract social representations form linguistic corpus. *Japanese Journal of Experimental Social Psychology*, *35*(3), 278–291.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. NY, NY: Farrar, Straus and Giroux.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Inc.

Longhurst, R. (1996). Education as a commodity: The political economy of the new further education. *Journal of Further and Higher Education*, *20*(2), 49–66.

Lowenberg, P. (2002). Assessing English proficiency in the expanding circle. *World Englishes*, *21*(2), 431–435.

Manalo, J., & Wolfe, E. (2000). *A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language* (Paper presented at the Annual Meeting of the American Educational Research Association (New

Orleans, LA, April 24-28, 2000) No. TM 031 430). American Educational Research Association.

Mauranen, A. (2003). Academic English as lingua franca - a corpus approach. *TESOL Quarterly*, *37*, 513–527.

Maycock, L., & Green, T. (2005). The effects on performance of computer familiarity and attitudes towards CB IELTS. *Research Notes, 12*, 4–8.

McClatchy, V., & Cooper, M. (1924). A psychological study of linguistic abilities with reference to the results of word association tests. *Journal of Experimental Psychology*, *7*(5), 371–381.

McCrum, R. (2010). *Globish: How the English language became the world's language*. New York, NY: Norton and Company.

McMurtry, J. (1991). Education and the market model. *Journal of the Philosophy of Education*, *25*(2), 209–217.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, *18*(4), 333–349. doi:10.1177/02655322010800402

McNamara, T., & Ryan, P. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. doi:10.1080/1543303.2011.565438

McPerren, A. (2007). The commodity market and university education. *University Business, 10*(10), 61.

Mead, M. (1926). The methodology of racial testing: Its significance for sociology. *American Journal of Sociology*, *31*(5), 657–667.

Mead, M. (1927). Educational research and statistics: Group intelligence tests among Italian children. *School and Society*, *25*(642), 465–468.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). NY, NY: MacMillan.

Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: Sage.

Naidoo, R. (2007). *Higher education as a global commodity: the perils and promises for developing countries*. London: The observatory on borderless higher education. Retrieved from www.obhe.ac.uk

Nelson, C. (1992). My language, your culture: Whose communicative competence? In B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 327–339). Chicago, IL: University of Illinois Press.

Nissan, S., DeVincenzi, S., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (Research Report No. 51). Monograph Series. Princeton, New Jersey: ETS.

Noble, D. (2003). Digital diploma mills. In B. Johnson, P. Kavanagh, & K. Mattson (Eds.), *Steal this university: the rise of the corporate university and the academic labor movement* (pp. 33–48). NY, NY: Rutledge.

Obst, D., & Forester, J. (2006). Perceptions of European higher education country report: USA. *Perception of European higher education in third countries*. Academic Cooperation Association Secretariat. Retrieved from http://www.acasecretariat.be/02projects/Perceptions.htm

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington D.C.: Office of Technology Assessment.

Ohanian, S. (2000). You say stakeholder: I say robber baron. *Language Arts*, *78*(2), 148–156.

Pelligrino, J. (2004). *The evolution of educational assessment: Considering the past and imagining the future* (Lecture Series). Princeton, New Jersey: ETS.

Pino-Silva, J. (2007). Student perceptions of computerized tests. *ELT Journal*, *62*(2), 148–156.

Popham, N. J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practices*, *16*(2), 9–13.

Powers, D., Kim, H.-J., Yu, F., Weng, V., & VanWinkle, W. (2009). *The TOEIC Speaking and writing tests: Relations to test taker perceptions of proficiency in English* (Research Report No. RR-09-18). Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-09-18.pdf

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing*, *14*(3), 304–314.

Ren, J., Bryan, K., Min, Y., & Wei, Y. (2007). Language preparation and the first year experience: What administrators and policy makers should know. *Florida Journal of Educational Administration and Policy*, *1*(1), 11–24.

Rosenfeld, M., Leung, S., & Oltman, P. (2003). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (Research Report No. MS - 21: RM-01-03). Princeton, NJ: ETS. Retrieved from http://www.ets.org/research/policy_research_reports/rm-01-03_toefl-ms-21

Sadighi, F., & Zare, S. (2006). Is listening comprehension influenced by the background knowledge of learners? A case study of Iranian EFL learners. *The Linguistics Journal*, *1*(3), 110–126.

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

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

Schonemann, P., & Heene, M. (2009). Predictive validities: Figures of merit or veils of deception? *Psychological Science Quarterly*, *51*(2), 195–215.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for English as a lingua franca. *International Journal of Applied Linguistics*, *11*(2), 133–158.

Seidlhofer, B. (2005). English as a lingua franca. *ELT Journal*, *59*(4), 339–341. doi:10.1093/elt/cci064

Seidman, I. (2006). *Interviewing as qualitative research* (3rd ed.). NY, NY: Teachers College Press.

Selvarajah, C. (2006). Cross-cultural study of Asian and European student perception: The need to understand the changing educational environment in New Zealand. *Cross-cultural Management: An International Journal*, *13*(2), 142–155. doi:10.1108/13527600610662320

Smith, J., Flowers, P., & Larkin, M. (2009). *Interpretative phenomenological analysis: theory, method and research*. Los Angeles, CA: Sage.

Smyth, T. S. (2008). Who is No Child Left Behind leaving behind? *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *81*(3), 133–137.

Spolsky, B. (1990). The prehistory of the TOEFL. *Language Testing, 7*(1), 98–118.

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

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, *100*(4), 961–976.

Stansfield, C. (1986). *A history of the test of written English: The developmental year* (Research Report). Princeton, New Jersey: ETS. Retrieved from http://www.eric.ed.gov/PDFS/ED275199.pdf

Stricker, L., & Attali, Y. (2010). *Test takers' attitudes about TOEFL iBT* (Research Report No. RR-10-02 TOEFLiBT-13). Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-10-02.pdf

Stricker, L., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (Research Report No. RR-05-12, TOEFL MS-32). Monograph Series. Princeton, New Jersey: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-12.pdf

Stricker, L., Wilder, G. S., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as Foreign Language. *Computers in Human Behavior*, *20*, 37–54.

Struyven, K., Dochy, F., & Janssens, S. (2002). Students' perceptions about assessment in higher education: A review. Presented at the Joint Northumbria/ Earli SIG Assessment and Evaluation Conference: Learning communities and assessment cultures, August 28-30, 2002, University of Northumbria as Newcastle.

Sunderman, G. (2006). *The unraveling of No Child Left Behind: How negotiated changes transform the law*. Cambridge, MA: The Civil Rights Project at Harvard University.

Templer, B. (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market. *Journal for Critical Education Policy Studies, 2*(1). Retrieved from http://www.jceps.com/?pageID=article&articleID=21

Triplett, C. F., & Barksdale, M. A. (2005). Third through sixth graders' perceptions of high-stakes testing. *Journal of Literacy Research, 37*(2), 237–260.

Tsai, Y., & Tsou, C.-H. (2009). A standardized English language proficiency test as the graduation benchmark: Student perspectives on its application in higher-education. *Assessment in Education: Principles, Policy and Practice*, *16*(3), 319–330.

University of New Mexico Division of Enrollment Management. (2011). *Official Enrollment Report Fall 2011* (Research Report). Official Enrollment Report. Albuquerque, NM: University of New Mexico. Retrieved from http://registrar.unm.edu/reports--statistics/fall2011oer.pdf

United States Department of Education. (2012). *President Obama: Our children can't wait for Congress to fix No Child Left Behind, announces flexibility in exchange for reform for ten states* (Press Release). Washington DC: US Department of Education. Retrieved from http://www.ed.gov/news/press-releases/president-obama-our-children-cant-wait-congress-fix-no-child-left-behind-announc

Van Manen, M. (1990). *Researching lived experiences*. Albany, NY: State University of New York Press.

Viks-Freibergs, V., & Freibergs, I. (1976). Free association norms in French and English: Inter-linguistic and intra-linguistic comparisons. *Canadian Journal of Psychology*, *30*(3), 123–133.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335–253.

Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Qualitative Linguistics*, *12*(2-3), 111–122. doi:10.1080/09296170500172403

Wildavsky, B. (2010). *The great brain race: How global universities are reshaping the world*. Princeton NJ: Princeton University Press.

Wilson, K. (1989). *Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experiences in the TOEIC testing context* (Research Report No. RR-89-39, TOEIC-RR-01). Princeton, NJ: ETS.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing*, *24*(2), 251–286. doi:10.1177/0265532207076365

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. Hornberger, (Eds.) *Encyclopedia of language and education (pp177-196)*. NY, NY: Springer Science and Business Media LLC.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. doi:10.1177/0265532209349465

Yu, G. (2007). Student's voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing*, *24*(4), 539–573. doi:10.1177/0265532207080780

Zareva, A. (2005). What is new in the new TOEFL - iBT 2006 test format? *Electronic Journal of Foreign Language Teaching*, *2*(2), 45–57.

Zhang, Y. (2008). *Repeat analyses for TOEFL iBT* (Research Report No. RM-08-05) (p. 17). Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-08-05.pdf