Correlations in the new TOEFL era: An investigation of the statistical relationships between iBT scores, placement test performance, and academic success of international students at Iowa State University

by

Marc Manganello

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English as a Second Language/Applied Linguistics (Language Assessment)

Program of Study Committee

Barbara Schwarte, Major Professor

Volker Hegelheimer

Marcia Rosenbusch

Iowa State University

Ames, Iowa

2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Possibly the single most fundamental concern in language assessment is the "usefulness" of test scores – how accurately and completely test scores communicate to users of those test scores (i.e., those who make decisions based on those scores) what they want to know about the language abilities of the individuals who have taken the test (Bachman & Palmer, 1996; Douglas, 2010). The usefulness of tests is particularly an issue with tests that play a role in high-stakes decisions like whether or not to admit an international student to a university. A particularly relevant example of a test associated with such high stakes would be the TOEFL (Test of English as a Foreign Language), which has in the past decade undergone a remarkable transition from its old paper-based version to a computerized version and finally to a more thoroughly modernized internet-based version. Ideally, as a result of the extensive research and long development process that went into the development of this new version of the TOEFL, universities should now be able to enjoy newfound faith in their usefulness as indicators of the language skills and knowledge they desire of applicants.

The TOEFL and its Role in University Admissions

International applicants to academic programs in the USA who are non-native speakers of English are nearly invariably required to demonstrate sufficient proficiency with English in order to be considered for admission. Evidence of this proficiency is usually provided in the form of a score on a major, standardized, norm-referenced test of academic English. Falling into this category include tests such as the recently launched Pearson Test of English, the academic version of the IELTS (International English Language Testing System), the MELAB (Michigan English Language Assessment Battery) and the TOEFL. Such tests are designed to, as holistically as possible, evaluate non-native English speakers' mastery of English in an academic setting (Alderson, Krahnke & Stansfield, 1987) and provide institutions with scores from which inferences may be made about applicants'

chances of success in a study program where coursework will be done in English (Chalhouh-Deville, 2003).

Of these tests, the TOEFL has remained the most widely recognized and trusted test of English used for admissions purposes worldwide since Educational Testing Service (ETS) began offering it in 1964 (Stevenson, 1987). Although the IELTS in particular has been making inroads on TOEFL's dominance in North America (Inside Higher Ed, 2008), the TOEFL remains the most commonly accepted test of academic English proficiency used for admissions purposes (Educational Testing Service, 2011a). Educational Testing Service (2011a) boasts on the TOEFL website that the TOEFL is currently accepted by more than 8000 institutions worldwide, including nearly every university in the USA and Canada, and is taken by nearly a million people each year.

Though widely relied upon, the TOEFL has long been subject to criticism and doubts over its usefulness for predicting academic success. This has been particularly true of the old, paper-based version of the TOEFL known as the PBT (paper-based test), which was the only version of the TOEFL that was available until the computer-based TOEFL became available in 2003. Studies during the 20th century repeatedly demonstrated scores on proficiency tests like the TOEFL PBT to be poor predictors of academic success (Graham, 1987; Simner, 1998; Chalhouh-Deville, 2003). Concerns also arose over the "washback" of the PBT version – its impact on the way EFL (English as a foreign language) is taught. The very high stakes associated with the TOEFL fostered the harmful attitude among both instructors and prospective students that mastery of the specific skills necessary to obtain higher PBT scores were more important than the development of speaking, writing, and pragmatic skills that learners would benefit from once admitted to a program (Hamp-Lyons, 1998).

It is also a tenet of language assessment that no one language test, no matter how carefully designed it may be or how powerful an argument may be presented for its validity, is a perfectly dependable indicator of the skills that it is meant to measure (Alderson, Krahnke & Stansfield, 1987; Bachman & Palmer, 1996). It has therefore become a common practice among American universities to administer an institutional test of English proficiency to international students upon their arrival (Douglas, 2003). Although it is assumed at this point that incoming international students possess at least the minimum

necessary level of proficiency in that they have achieved the required score on a test such as the TOEFL, these institutional tests (or series of tests) are intended to further gauge these students' mastery of skills such as reading, listening, and writing. Ideally, the results of these tests help universities ensure that non-native speakers of English are placed into English classes that are appropriate for their ability level and best suited to their outstanding English proficiency needs (Douglas, 2003).

The Internet-Based TOEFL

In 2005, the implementation of a new, internet-based version of the TOEFL, the iBT (internet-based test), represented the culmination of more than a decade of research and effort on the part of ETS to develop a modernized, computerized version of the TOEFL that would better suit the needs and expectations of the institutions that rely on TOEFL scores for admission decisions (Taylor & Angelis, 2008). Inclusion of assessment of language production (the ability to actually communicate using English) in the test was a major goal in these efforts (Taylor & Angelis, 2008). Although the PBT is still considered no less official or valid a version of the TOEFL, and PBT scores are still accepted by most institutions, iBT scores now constitute the vast majority of TOEFL scores sent to institutions.

Because the iBT scores reflect production of language (writing and speaking) and tasks requiring the integrated use of multiple skills, TOEFL scores should now be a better indicator of the extent to which international applicants posses the practical capabilities as well as knowledge that are conducive to success. However, can connection between TOEFL iBT scores and academic performance in English be statistically demonstrated in addition to presumed to possibly exist? While the shortcomings of the PBT were much discussed during the 20th century, and the lack of a clear relationship between academic success and PBT scores have been well documented, the predictive strength of the relatively new iBT as well as its impact on EFL instruction has not yet been thoroughly explored.

Purpose of This Study

This study shall examine how dependably international students' performance on the TOEFL iBT predicts their university academic success. To this end, this study shall test the strength of the statistical relationships between international students' iBT scores, scores on Iowa State University's own placement test, the EPT (English Placement Test), and grades obtained in English classes. At least some relationship between different tests of English proficiency can be realistically expected since the tests are designed to measure similar "constructs" – the knowledge or skills that tests are designed to evaluate (Douglas, 2010). If examinees have the necessary combination of education, practical experience, and aptitude for second language acquisition to perform well on a test of English proficiency such as the TOEFL, it follows logically that these same examinees should stand the best chances of performing well on other tests of English proficiency.

The key issue at hand is the strength of the statistical relationships. Tellingly strong relationships between scores on the iBT and performance on the EPT could be considered evidence that the administration of the EPT in its current form may be redundant when an iBT score has already been reported. Meanwhile, strong relationships between iBT scores and the performance of international students in classes would serve as evidence that iBT scores accurately reflect the degree to which international applicants have mastered the comprehension and communication skills that are conducive to academic success. However, only fair or weak relationships would weaken the case for the use of iBT scores for class placement purposes and also cast doubt on the usefulness of increasing minimum required TOEFL scores (or even of maintaining such a requirement in the first place) as a way to better select only the best qualified international applicants.

Anticipated Results

As the listening and reading sections of both the TOEFL iBT and ISU's EPT are similarly designed and intended to test similar constructs, one would expect a high

correlation to be noted between the two tests in light of the vigorous validation processes that was part of the development of both. The strong documented correlations between the TOEFL and other language tests during the PBT era (Stevenson, 1987; Graham, 1987) would presumably hold true for the iBT as well. Furthermore, the changes to the TOEFL during the transition from PBT to iBT should theoretically enable the iBT to assess proficiency in a way that better models what is actually expected on students in a university classroom. As a result, this study should uncover far stronger relationships between iBT scores and academic success in language-intensive coursework than the weak relationships between PBT scores and academic performance that were documented during the 20th century. If the analysis of the data should fail to uncover a more significant correlation than was typical of the PBT era, however, the findings of this study should nevertheless be useful to the university for the review of current EPT administration and class placement procedures.

CHAPTER 2
BACKGROUND AND THEORETICAL CONSIDERATIONS

This chapter provides a theoretical context for this study of the relationships between iBT scores and other variables. The rationale for English proficiency testing and the development of the PBT version of the TOEFL are discussed, as well as the understanding of English "proficiency" according to which the PBT was originally designed. Issues pertaining to the usefulness of PBT scores are identified, as well as the significance of the changes to the TOEFL made by ETS during the development of the iBT. The role of institutional tests, such as Iowa State University's EPT, are also discussed. Finally, the issues pertaining to proficiency and placement testing that are relevant to this study are summarized.

Proficiency Testing

The administration of the kinds of tests described as "proficiency tests" began in the 20th century in response to a growing number of international applicants to universities in English-speaking countries in the 1950s. By the 1960s, the persistence and strengthening of this trend led to demand for standardized, reliable tests that could be used to evaluate English proficiency for admissions purposes (Taylor & Angelis, 2008). This occasioned the development of the CELT (the forerunner to the IELTS) in the UK and the TOEFL in the USA. By the 1970s, nearly all academic programs in the English-speaking world had begun to require that non-native English speaking international applicants submit a score on a proficiency test in order to be considered for admission. It remains the case more than a decade into the 21st century that institutions nearly invariably require prospective international students to submit some form of evidence of their ability to use and understand English in an academic setting. Scores on major proficiency tests like the TOEFL have been the most commonly accepted form of evidence of proficiency (Simner, 1998). Douglas (2003) describes proficiency tests under the heading of "admissions testing" due to their role in admissions policies and outlines the rationale behind such tests:

> Proficiency tests are not based on any specific course of study (as are achievement tests, for example) but are intended to measure the ability to use English in specific situations in which the learners will find themselves in the future, regardless of the circumstances in which they acquired the language. Since they are often taken months before a prospective student arrives on campus, proficiency tests are therefore often used to look forward, aiming to help make predictions about the probability that a particular applicant will be able to cope with the demands for English language use in the context of college and university study (p. 3).

The purpose of proficiency tests is therefore to help universities identify international applicants who are sufficiently prepared for coursework in an English-speaking environment. Ideally, scores on these tests should help the institutions that rely on them to make admission decisions ensure that only international applicants with reasonably high prospects of success are admitted to their academic programs.

The PBT Version of the TOEFL

The Modern Language Association's Center for Applied Linguistics began work on the development of the TOEFL in 1961. Several other US organizations, including the Institute of International Education and the National Association of Foreign Student Advisors were involved in the project as well (Taylor and Angelis, 2008). The first major step in designing a new proficiency test was "to attempt to identify a common core of language abilities that would be relevant to the range of situations in which students would find themselves at the university" (Taylor and Angelis, 2008, p. 29). The TOEFL was originally designed according to "the ability approach to language teaching" in which language was seen as "composed of separately definable components such as a sound system, grammar, and vocabulary" (p. 29-30). This understanding of language competence, which was influenced by Chomsky's emphasis on the structure of language (Douglas, 2010), would be repudiated in the following decades, when the idea of language proficiency as

communicative competence increasingly came to center stage. At the time, however, the following list of abilities drafted by psychologist John Bissell Carroll found wide support as the potential basis for an English proficiency test (Taylor & Angelis, 2008, p.29):

1. Knowledge of structure.
2. Knowledge of general-usage lexicon.
3. Auditory discrimination (of phonemes, allophones, and suprasegmentals.)
4. Oral production (of phonemes, allophones, and suprasegmentals.)
5. Reading (in the sense of converting printed symbols to sound.)
6. Writing (in the sense of converting sound to printed symbols, i.e., spelling.)
7. Rate and accuracy of listening comprehension.
8. Rate and quality of speaking.
9. Rate and accuracy of reading comprehension.
10. Rate and accuracy of written composition.

Considering that assessment of any actual testing of language production went on to be so conspicuously (and controversially) missing from the PBT version of the TOEFL, it is interesting to note here that writing and speaking are mentioned on Carroll's list (though with emphasis on "correctness" and "accuracy" rather than effective communication). However, incorporating assessment of writing and speaking into the test was a problem for which there was no practical and expedient solution at the time. Taylor and Angelis (2008) note that during the development of the TOEFL "the difficult areas proved to be speaking and writing, numbers 8 and 10, respectively, on Carroll's list" (p. 30). Ultimately, speaking and writing were omitted from the 140 multiple-choice item format of the PBT that ETS eventually settled on. In adopting the purely multiple-choice format, the designers of the PBT borrowed heavily from the format of other language tests in use at the time, such as the American University Language Center Test (Taylor and Angelis, 2008). When the PBT was officially launched in 1964, its format and the idea of proficiency that it was designed to test therefore represented attitudes towards language testing, language learning, and linguistics in general that were well-established in the 1960s. While attitudes would change considerably during the rest of the 20th century, the TOEFL however remained fundamentally the same.

Content of the PBT

The PBT, the original, paper-based version of the TOEFL, which ETS began to offer in 1964, uses the same multiple-choice format that has been characteristic of other ETS tests such as the SAT (Scholastic Aptitude Test) and GRE (Graduate Record Examination). PBT test items consist of 140 multiple-choice questions in three categories: listening comprehension (50 questions), structure and written expression (40 questions), and reading comprehension (50 questions).

In the listening comprehension section of the PBT, examinees listen to recorded dialogues or monologues in English and after each are asked by a narrator to indicate the best of four possible answers to each item. The structure and written expression section of the PBT features multiple-choice cloze items, where examinees must choose a response that, when inserted into the blank, best completes a written sentence, and items where examinees must identify which of four underlined words would need to be changed in order for a sample sentence to be grammatically correct. In the reading comprehension section, examinees choose the best answers to questions based on printed texts.

Raw scores on all three sections of the PBT are converted to a scale ranging from 31-67 for sections 1 (listening comprehension) and 3 (reading comprehension), while the scale ranges from 31-68 for section 3 (structure and written expression). The scores for all three sections are averaged and multiplied by 10 for a total score range of 301-677 (Educational Testing Service, 2011c). An essay test called the TWE (Test of Written English), when it is administered as part of the PBT, is scored separately on a scale of 0-6. Examinees' scores on the TWE have no bearing on their overall PBT scores; only the PBT scores have served as the basis for admission to universities.

Justification of the PBT

As has been noted previously, the multiple-choice format of the PBT version of the TOEFL precluded the possibility of sections of the test assessing an examinee's production of language. While the inclusion of assessment of language production would go on to become a major goal in the development of a new version of the TOEFL, there were some advantages to the multiple-choice format of the PBT. According to Livingston (2009):

> The multiple-choice format has come to dominate large-scale testing, and there are good reasons for its dominance. A test-taker can answer a large number of multiple-choice questions in a limited amount of testing time. The large number of questions makes it possible to test a broad range of content and provides a good sample of the test taker's knowledge, reducing the effect of "the luck of the draw" (in the selection of questions) on the test taker's score. The responses can be scored by machine, making the scoring process fast and inexpensive, with no room for differences of opinion (p. 1-2).

As Livingston (2009) notes, one of the strongest arguments that can be put forward in favor of multiple choice testing is that it is possible for examinees to answer a larger number of questions in a given period of time. The increased number of questions increases the statistical reliability of the test, minimizing the margin of error and increasing the likelihood that the final score will be an accurate representation of a test taker's ability level relative to other test takers. The "luck of the draw" mentioned by Livingston could particularly be an issue with essays tests. Due to the amount of time necessary to complete an essay task, it may not be possible for examinees to compose more than one or two essays. Examinees' level of familiarity with the one or two essay topics they are given could significantly affect their performance for reasons other than their level of English proficiency. Another major advantage of multiple-choice testing in comparison with essays is that tests can be quickly machine scored, which also circumvents any potential problems with variance due to rater biases, which can be problematic with essays (Livingston, 2009).

Douglas (2010) claims that multiple-choice test items "allow test takers to demonstrate their ability to control very fine distinctions in vocabulary, grammatical structures, phonology, or comprehension of content, but they are notoriously difficult to develop" (p. 50). Test items in which examinees must select the correct response can be very challenging when they are well-designed and are a proven method of assessing knowledge and comprehension. The challenge that ETS faced for the rest of the 20th century was to design and implement a test that would assess the practical use of language as well as

knowledge, be practical to administer to a large number of test takers, and assure consistency in scoring.

Concerns Over the Role of the PBT in University Admissions

TOEFL validation efforts strive primarily to verify that the test does indeed accurately and reliably measure proficiency with English – a formidable task when something as multifarious as the whole notion of proficiency is difficult to define and no one definition is ever set in stone. However, another important question concerning the usefulness of the TOEFL to the institutions that have made use of it for admissions purposes is whether inferences about an international applicants' readiness to begin a study program can indeed be accurately made based on TOEFL scores. At the heart of the controversy is the fact that language proficiency alone is not necessarily indicative of academic aptitude and potential for success (Graham, 1987; Simner 1998). It does follow logically that classes will be more difficult for students who have trouble understanding teachers and texts and they may not always be able to express themselves clearly. International applicants who are successful enough at learning EFL to obtain a target TOEFL score thereby demonstrate some degree of the necessary ambition and study skills as well. On the other hand, there are a great many other factors besides ability to understand and communicate in English that contribute to whether or not an international student is ultimately successful in his or her chosen course of study (Simner, 1998).

Another topic related to the role of the TOEFL in admissions that is discussed in this section of Chapter 2 is the washback associated with the PBT. Although concrete documentation of the washback that occurred during the PBT era is scarce (Bailey, 1999), the practice of "teaching to the test" seemed likely to occur when the stakes are as high as they tend to be for learners who are taking the TOEFL (Hamp-Lyons, 1998; Bailey, 1999). The PBT's important role in university admissions in the 20th century fueled concerns that EFL teaching focused too much on comprehension and knowledge of structure as a result. Skills that were not tested by the PBT, such as writing and speaking, were less likely to be

considered important learning objectives despite their importance for success once learners obtain admission to a university and begin coursework (Wall and Horák, 2006).

Correlation of TOEFL Scores with Academic Success

During the PBT era, many studies investigated the statistical relationship between scores on the PBT and indicators of academic performance such as GPA. The relationships that were observed in these studies varied in strength, but in most cases they were not very convincing and failed to support the idea that English proficiency was essential for success (Graham, 1987). During correlation studies, a correlation coefficient of more than 0.70 is necessary in order to claim that one variable is at least 50% determined by the same factors that determine the other. Studies correlating TOEFL scores with academic performance tended to fall far short of that (Graham, 1987; Simner, 1998). Several studies in which only a marginally positive statistical relationship between PBT scores and other indicators of academic success was found are identified in Table 1.

Table 1: Studies cited by Graham (1987) and Simner (1998)

Researcher(s)	Year	Variable Correlated with PBT Score	Correlation Coefficient
Sugimoto	1966	Successful Completion of Program	-0.046
Hwang and Dizney	1970	Graduate GPA	0.19
Sharon	1972	Graduate GPA	0.26
Light, Xu, and Massop	1987	Graduate GPA	0.14
Hughey and Henson	1993	Undergraduate GPA	0.19

There is good reason not to expect extremely strong relationships between English proficiency test scores and grades earned in classes when even native English-speaking students (all of whom would theoretically score very high on tests such as the TOEFL) obtain a wide range of grades. But the slightly negative correlation coefficient documented in the

Sugimoto (1966, cited in Graham, 1987) study as well as the only marginal positive coefficients observed in the other studies noted above do not support the theory that English proficiency is a factor in determining success, either in the form of GPA or successful completion of a program. Simner (1998) remarks of the above studies:

> It is worth noting that the TOEFL scores in these investigations ranged from approximately the 5th to the 99th percentile. Therefore, it is unlikely that these low level correlations could have resulted from a restricted TOEFL range. Instead, it would seem that the magnitude of these correlations reflect a genuine lack of any meaningful relationship between the TOEFL and academic achievement (p. 262).

Simner (1998) reports a similar lack of a significant positive relationship when completion of an academic program rather than grades serves as the definition of academic success:

> The findings, which are available from the University of Western Ontario (Simner, 1995), revealed that undergraduate students with TOEFL scores in the 550–579 range not only performed as well as students with scores in the 580–677 range but also as well as their Canadian counterparts. For example, among the students registered in the Faculty of Social Science, the graduation rate was 83% for those with TOEFL scores from 580–677, for those with TOEFL scores between 550–579 the graduation rate was 82% and for the Canadian students it was 84% (p. 263).

Such findings were not necessarily indicative of flaws in the PBT itself because, unlike other ETS-designed tests like the SAT and GRE, the TOEFL is not intended to measure academic aptitude (Graham 1987, Simner 19980. Such research did however fuel discussion of the overreaching purpose of the TOEFL (Taylor & Angelis, 2008) and prompted renewed efforts on the part of ETS to clarify appropriate and intended use of TOEFL scores by institutions (Simner, 1998). In particular, PBT user manuals provided by ETS specifically warned client institutions against the use of applicants' TOEFL scores as a sole criterion for admission (Simner, 1998; Stevenson, 1987).

Simner (1998) contends that the nevertheless highly prevalent practice among universities in Canada and the USA of requiring a minimum TOEFL score for admission has

been unnecessarily blocking the admission of otherwise qualified international students who may have been successful. Furthermore, Simner (1998) noted a trend among universities to increase their minimum acceptable TOEFL scores rather than give higher regard to other criteria that had been demonstrated to be more accurate predictors of success:

> In Ontario, for example, between 1991 and 1995 three universities (Guelph, Toronto, and Western) raised their undergraduate cutoffs from 550 to 580. In fact, by 1995, of the 18 Ontario universities that made use of the TOEFL, ten universities (Carleton, Guelph, McMaster, Ottawa, Queens, Ryerson, Toronto, Waterloo, Western, and York) had minimum cutoffs that ranged from 580 through 600 (Byrne, 1995). The picture is very similar throughout the rest of Canada. For instance, according to their 1996/97 academic calendars the Universities of Alberta, Calgary, Dalhousie, Regina, and Simon Fraser all required minimum scores that ranged from 580 through 600. Because a TOEFL score of 580 is equivalent to the 83rd percentile while scores in the vicinity 600 are near the 90th percentile, these increases mean that substantial numbers of otherwise qualified nonnative English speaking applicants could be denied admission to these universities (p. 263).

The still unresolved debate over the role of English proficiency as an admissions requirement proceeds under the assumption that accurate inferences about English proficiency can be made on the basis of scores on proficiency tests. Strong correlations between TOEFL scores and other tests of English proficiency, in addition to the documented statistical reliability of the TOEFL, served as some assurance that it was (Stevenson, 1987). Nevertheless, the exclusion of any assessment of language production in favor of testing knowledge of "discrete points" concerning the grammatical structure of the English language, isolated from any particular context, fueled doubts that skills tested by the PBT were the kinds of skills that would be most useful to international students during actual coursework (Hamp-Lyons, 1998).

Washback and its Effect on the Usefulness of Scores

Further cause for concern was the washback of the PBT: its impact on learner attitudes and EFL teaching practices The very high stakes associated with the TOEFL led to concern that PBT preparation methods and EFL instruction in general was becoming too concerned with teaching the knowledge of structure necessary to obtain a better PBT score at the expense of emphasis on useful skills such as writing and speaking.

Washback can occur in the form of impact on individuals (test-takers, students, and teachers) or impact on society and education systems (Bachman & Palmer, 1996). Individual learners in particular can be impacted negatively from the stress while preparing to take a high stakes test, in some cases leading them to neglect other commitments such as attending classes. Failure to pass a very important test (or obtain a sufficiently high score if there is no real "passing" score) may take a severe emotional toll on test-takers as well (Karabulut, 2007). The test fee (usually in excess of US $100) can also represent a serious financial hardship to test-takers in some countries (Bailey, 1999; Chalhouh-Deville, 2003). Individual teachers may also succumb to pressure to change their teaching objectives and behave differently than they would normally consider to be in their learners' best interests (Hamp-Lyons & Shohamy, 2003; Karabulut, 2007). For the most part, however, the controversy involving the washback of the TOEFL (and particularly the PBT version of the TOEFL) centered on its potential impact on whole education systems and attitudes towards EFL abroad. It is generally understood and agreed upon in the field of language assessment that some degree of impact on educational systems is inevitable when the stakes associated with a test are as high as they are in the case of the TOEFL (Bachman, 1996; Bailey, 1999).

Washback is not always harmful, however. EFL/ESL programs can be positively impacted by large-scale, high-stakes proficiency tests if the goal of better preparing learners for these tests results in higher standards, clearer learning goals, and increased consensus on objectives that facilitates the establishment of institutional or national policies (Hamp-Lyons, 1997). However, since the 1990s most discussion of washback has centered on the known or perceived negative impact of high-stakes tests (Hampy-Lyons, 1997). When whole programs seek to accommodate the best interests and desires of learners whose future plans are so

contingent on obtaining a sufficiently high test score, teaching to this test may occur in a way that compromises the overall quality of an EFL or ESL program.

If everything seems to depend on TOEFL scores, learners themselves may demonstrate a lack of interest in any topics or material that they do not think pertain directly enough to performance on the TOEFL (Taylor & Angelis, 2008; Wall & Horák, 2006). Because the PBT does not assess the production of language aside from the Test of Written English, learners demonstrated disinterest in the speaking and writing skills that would be necessary success in their coursework. The inclusion of the TWE was applauded as a step in the right direction to counteract this trend, according to Taylor & Angelis (2008):

> The introduction of the TWE was met with much support from the language teaching community because the test required actual writing. Not only was this seen as a sign of a move toward testing of real language abilities, but also as support for the teaching of writing. Particularly in ESL settings such as intensive English programs in the United States, teachers had long bemoaned the fact that their efforts to teach writing had met with little enthusiasm from students because the students knew that the test they would take as part of their pursuit of university admission was a completely multiple-choice test (p. 35).

However, as the TWE is scored separately from the rest of the PBT, a high score on the essay does not contribute to a higher score towards the minimum set by an institution, marginalizing the importance of this part of the test. In some cases when the PBT is administered institutionally, the Test of Written English is not even included. It followed therefore, unfortunately, but logically, that the production of language in EFL instruction continued to be downplayed. Wall and Horák (2006) note during their observations of TOEFL preparation classes in various countries in Eastern Europe that "many of the teachers and students felt that speaking was not an important skill to practice or learn because it was not going to be tested" (p. 70). Meanwhile, Hamp-Lyons (1998) contended that teaching to the test – concentration on the discrete skills that are tested by the PBT – occurred to such a degree during the PBT era that the scores of learners from some countries were widely held to be 20-30 points higher than accurately reflected an examinee's overall English

proficiency. Hamp-Lyons (1998) claimed this strategic yet unethical approach to TOEFL preparation, where increasing scores is the only objective, manifested itself particularly visibly in many of the textbooks used in TOEFL preparation courses around the world during the PBT era:

> Because the books are built around the model of the test and because the test is not intended to reveal or reflect a model of language in use, even if it is built upon one, teacher and learners find themselves teaching - and trying to learn - discrete chunks of language rules and vocabulary items without context or even much co-text (p. 332).

If TOEFL preparation practices and materials during the PBT era were indeed so focused on structural issues rather than expression and communication because of overwhelming pressure to increase PBT scores, it is not surprising that PBT scores tended not to reflect many of the skills that constitute language proficiency. Moreover, the usefulness of TOEFL scores as indicators of potential for success was compromised as a result.

Incorporation of Language Production into the TOEFL

Although the PBT version of the TOEFL has received many favorable reviews in addition to criticism, with Stevenson (1987) describing the TOEFL as "best of its breed" at the time despite its perceived shortcomings (p. 81), many of the issues raised by critics were already topics in ETS' own self-determined agenda for an improved TOEFL. According to Taylor and Angelis (2008):

> By the 1980s, Carroll's 1961 idea of integrative language ability had been expanded and explored by researchers in applied linguistics using such terms as communicative competence. The TOEFL program attempted to gather information and conduct analyses that would help to determine how communicative competence could be measured by the TOEFL (p. 34).

Discussion of new tasks to better test the communicative competence of examinees led to the aforementioned decision to reinstate the TWE after it had originally been excluded from the PBT due to practical concerns. A speaking test called the TSE (Test of Spoken English) was also developed to serve as another language-production adjunct to the PBT. Although the TWE went on to become a part of the standard PBT, it still was not always administered along with the PBT in all situations (Taylor & Angelis, 2008). Meanwhile, the decision was made to only include the TSE in special situations where the assessment of spoken language would be particularly important, such as for the evaluation of graduate teaching assistants (Taylor & Angelis, 2008).

Computerization of the test was seen as one way to facilitate the better integration of speaking and writing into the TOEFL (Enright et al., 2008). The first computerized version of the TOEFL, the CBT (computer based test) that was introduced in 1998, still consisted of very simple tasks and was in effect a computerized version of the PBT (Bannerjee, 2003). A significant development was that CBT scores were calculated according to a new scale into which the score on the written essay figured (Wang, Eignor, & Enright, 2008). However, there was still no assessment of speaking in the CBT (Bannerjee, 2003). The implementation of a completely new scoring system into which all four "modalities" (reading, listening, writing, speaking) factored was not realized until the iBT was launched in 2005.

The iBT Compared with the PBT

The iBT version of the TOEFL consists of four sections, each pertaining to a specific skill or "modality": reading, listening, speaking, and writing. Scores on each section range from 0 to 30. Scores on all four sections are combined into a "composite" iBT score that ranges from 0 to 120 (Educational Testing Service, 2011c). The reading and listening sections of the iBT have not changed dramatically compared to how they appear on the PBT. Multiple-choice items are still present in these sections, though computerization has allowed for some new features such as drag-and-drop matching items and a glossary to assist learners in the reading section (Educational Testing Service, 2011c). The most important difference between the iBT and PBT versions of the TOEFL is that speaking is now assessed as a standard part of the test, finally realizing this aspect of the original agenda for the TOEFL,

while writing performance is also fully integrated into the overall TOEFL score without constituting a separate, attached test. This means that scores for the production of language are now factored into the composite TOEFL score that is used for admissions decisions, as was so controversially not the case with the PBT.

The presence of "integrated" tasks in the writing and speaking sections of the TOEFL is a noteworthy development as well. During integrated tasks in the writing section, examinees read a passage and listen to a lecture, then have 15 minutes to compose a response to a question based on information in the text and lecture (Educational Testing Service, 2011c). This activity therefore tests the integrative use of language skills, since listening and reading comprehension factor into an examinee's success on the writing task. Integrated tasks in the speaking section function similarly – using a microphone, examinees record their responses to four questions based on information provided in the form of textual and/or audio input (Educational Testing Service, 2011c).

Potential Impact of the iBT

One of ETS' goals in developing a modernized version of the TOEFL was to positively impact TOEFL preparation practices and English teaching in general. If ETS has been successful, EFL and ESL practices should have adapted to reflect the new emphasis on communicative competence and integrative use of skills. As a result, TOEFL preparation should now result in prospective international students who are not only able to obtain higher TOEFL scores but are also better prepared to participate in an English-speaking environment once they are admitted to a program. More importantly, due to the emphasis on language production as well as understanding and the presence of integrative tasks on the test, the iBT should now more accurately measure overall language proficiency in a way that reflects how the language will be used during real university coursework. Theoretically, the transition from PBT to iBT should result in stronger correlations between TOEFL scores and real academic success, though questions remain concerning the extent to which English proficiency plays a role in academic success even when it is accurately represented by tests scores.

Administration of the English Placement Test at Iowa State University

Even when applicants are required to obtain a sufficiently high score on a proficiency test before being admitted to a university, it remains a common practice that universities administer placement tests of their own to newly-arrived international students (Douglas, 2003). Although ETS does maintain that making class placement decisions based on TOEFL scores constitutes appropriate use of the TOEFL (Simner, 1998), the TOEFL is designed to assess a wide variety of proficiency levels from very low to very high (Brown, 2003). Universities therefore find it expedient either to develop their own tests or purchase tests designed more specifically to separate learners within the intermediate to high range of proficiency levels that typically characterizes international students seeking to attend North American universities, per Brown (2003), who explains in further detail:

> Placement tests are designed for the population of students already at a particular institution, or just arriving, and they measure students' abilities in a particular language relative to the abilities of all other students at the institution... Such placement tests are administered for deciding what level of study is appropriate for each student, while in some cases they are used to decide which level of a series of integrated skills courses the students should take (p. 43).

Iowa State University's placement test, the EPT (English Placement Test) consists of three sections: writing, reading, and listening. The writing section determines which English composition class incoming international students are to be placed into, while the reading and listening sections determine if the same students need to be placed in special ESL classes focused on reading or listening skills.

EPT Reading and Listening Tests

The EPT reading and listening tests are 30-item multiple-choice/selected response tests similar in format to the PBT, in which answers are recorded on an answer sheet that is machine-scored afterward. Responses to the reading test items are in based on short printed texts, while responses to listening test items are based on speech in a video clip, whether

narration of the video or a recorded lecture. For both the reading and listening tests, 13 or more correct responses to the 30 items are necessary to "pass". Examinees who score 12 or lower on the reading test must take English 99R, and ESL class focused on development of reading skills. Those who score 12 or lower on the listening test must take English 99L, an ESL class focused on development of listening skills. Examinees who fail to obtain a passing score on either test must take both English 99R and 99L.

EPT Writing Test and Class Placement

The outcome of the EPT writing test, in which examinees write one essay on a given topic, determines placement into English classes focused on written communication. Incoming International undergraduate students are placed either into English 101B, 101C, or 150. English 101B, the lowest placement level, is a "review of English grammar in the context of writing" for international students who have exhibit difficulties with grammar that interfere with effective written communication (Iowa State University, 2011a). International students who demonstrate better mastery of grammar but are judged to still need help with writing in academic English are placed into English 101C, an ESL class that "prepares students for ENGL 150 and 250 and for writing in other disciplines"(Iowa State University, 2011a). International undergraduate students whose writing is judged sufficient to "pass" the writing test are placed into English 150, a basic non-ESL composition and communication course that is required of all ISU students.

International graduate students who pass the writing test are not required to take English 150 and are not placed in a composition class. Those who do not pass but do not need to be placed in English 101B are placed in English 101D, a graduate-level ESL composition class. English 101D focuses on "instruction in writing professional communication, academic papers and reports and in using published source material in writing" (Iowa State University, 2011a).

Although English 101 and 150 are primarily writing classes, coursework often involves speaking and presentations (especially in the case of English 101D and English 150). Nevertheless, placement in these classes is decided wholly by the EPT writing test. The results of the reading and listening tests have no bearing on composition class placement

The Current Role of the TOEFL in Class Placement

Incoming international students may be exempted from the EPT based on sufficient evidence of English proficiency beyond what is required for admission. Very high TOEFL scores (105 or more on the iBT or 640 or more on the PBT) are one way an incoming international student may be exempted from the EPT (Iowa State University, 2011b). Exempted students are not required to enroll in any English 99 classes and are placed automatically in English 150 (if undergraduate) or not required to take an English composition class (if graduate). A high writing score on the TOEFL is not sufficent to obtain exemption unless the composite score is the required 105, however.

Summary of Issues

In this chapter, three main issues with the use of the TOEFL for admissions purposes during the PBT era were identified:

1. Admission decisions based on TOEFL scores assume that English proficiency will play a major role in determining a student's success or failure. That English proficiency really does play such an important role, however, could not be demonstrated.
2. Practical concerns resulted in the exclusion of assessment of language production and integrative tasks from the PBT. As a result, the PBT's usefulness as a meaningful test of "proficiency" in the sense of ability to communicate and participate effectively in an English-speaking academic environment remained questionable.
3. The tremendous importance of PBT scores to the future educational and career plans of examinees led to an emphasis on structural issues such as grammar and vocabulary. This emphasis on knowledge alone rather than communicative competence detracted further from the likelihood that a PBT score reflected the

skills that would best qualify an international applicant to succeed in coursework at a university in an English-speaking country.

The following issues, though not pertaining the PBT version of the TOEFL specifically, concern global proficiency testing and its relationship to institutional placement testing:

1. According to ETS, appropriate use of TOEFL score would include class placement decisions based on scores. Though there would still be potential problems with the fact that the TOEFL measures a broader range of ability levels than most universities find useful for placement purposes, scores on the reading, listening, and writing sections of the iBT could potentially fill the same role as some institutional tests like the EPT if they prove to correlate strongly with sections of institutional tests designed to test the same skills.
2. Iowa State University does already use the TOEFL for placement purposes insofar as applicants with very high TOEFL scores are exempted from the EPT. A strong correlation between iBT composite scores and class placement could serve as justification of this policy, while a weak correlation may suggest that this policy should be reviewed.

CHAPTER 3
RESEARCH QUESTIONS AND METHODOLOGY

The same way the 20th century studies cited by Graham (1987) and Simner (1998) used correlations to ascertain the strength of the statistical relationship between PBT scores and academic success, this thesis study shall correlate iBT scores with international students' performance in other capacities where they are required to make use of their English proficiency. Using data on international students who have attended Iowa State University from the Fall 2009 to the Spring 2011 terms, the strength of the statistical relationship between iBT scores, performance on ISU's EPT, and grades obtained in English composition classes, will be investigated. The strength or weakness of these relationships that is revealed in this study will reflect on the usefulness of iBT scores to ISU, both in terms of their current role for admissions purposes and possible exemption from the EPT, and also the potential expansion of their role in determining class placement.

This chapter will begin with the identification of the specific research questions that this study shall pursue, followed by discussion of the theoretical issues identified in Chapter 2 to which each research question pertains. The data used in this study will then be identified and the statistical procedures that will be used based on these data will be outlined. After further explanation of the rationale behind the procedures that will be used to analyze data in this study, tables will display a summary of all of the correlations run in order to answer each research question. Finally, there will be discussion of the "truncated sample" of iBT scores that the participants in this study represent and acknowledgement that statistical relationships like those examined are insufficient to prove causal relationships between variables.

Research Questions

Collection and analysis of data in this study will proceed within the context of the following three research questions:

1. How strong is the statistical relationship between the iBT scores of incoming international students and the same individuals' performance on ISU's EPT?
2. How strong is the statistical relationship between international students' iBT composite scores and the grades they obtain in the English composition classes into which they are placed?
3. How strong is the statistical relationship between international students' scores on the four different sections of the iBT (reading, listening, speaking, and writing) and the grades they obtain in the English composition classes into which they are placed?

Theoretical Rationale for Research Questions

The correlation of iBT scores with EPT performance in order to obtain answers for research question #1 will assess the extent to which the iBT and ISU's EPT measure the same skills. There is no evaluation of speaking in the EPT, but the reading, listening, and writing tests that are part of the EPT assess skills that are also tested by the reading, listening, and writing sections of the iBT. This study will therefore test the strength of the relationship between performance on the reading, listening, and writing sections of both tests. If a strong relationship is noted between scores on the reading and listening sections of the iBT and the reading and listening tests that are part of the EPT, a case could be made for the use of iBT scores to decide if incoming international students would benefit from being placed in English 99R or 99L. Similarly, if a strong relationship between scores on the iBT writing section and placement into English composition classes based on the EPT writing test is observed, the administration of the EPT writing test may be redundant if the same skills have already been sufficiently tested by the iBT for an accurate placement decision to be made. Meanwhile, a strong relationship between iBT composite scores and class placement based on the EPT writing test would serve as justification for the Department of English's current policy where incoming international students with iBT scores of 105 or higher are exempted from the EPT.

Research questions #2 and #3 pertain to the greater overall issues of the role of the TOEFL in admissions decisions and the accuracy with which iBT scores forecast success in university coursework. There are two main reasons why grades in English composition

classes will serve as the variable representing academic success during correlations with iBT scores in this study. Most primarily, English classes are language intensive and English proficiency should therefore be of the greatest possible relevance to international students' performance in these classes. Furthermore, although English 150, 101B, 101C, and 101D are referred to as "composition classes" (as opposed to ESL classes like English 99L and 99R that focus on comprehension skills), and grades in these classes are determined more than anything else by performance on written assignments, oral communication as well as listening and reading comprehension are also important and necessary for success these classes. English composition classes therefore represent an academic environment where the integrative use of all four skills (reading, listening, writing, speaking) now tested by the iBT will be particularly important.

The answers to research question #2, provided by the correlation of iBT composite scores with grades in English composition classes, will reveal if there is a consistent pattern where international students with relatively high iBT scores also tend to obtain the highest grades in these classes. Strong relationships between grades and composite iBT scores would serve as evidence that scores on the iBT version of the TOEFL reflect mastery of English language skills that are important for academic success. Since these findings will be based on recent ISU data, they will be especially pertinent to the discussion of ISU's own admissions policy where an iBT composite score of 71 or higher is currently required for consideration for admission.

Finally, the answers to research question #3 will offer a more in-depth look at the relationship between iBT scores and academic performance when scores on specific sections of the iBT are correlated with grades in English composition classes. The results of these correlations should reveal if any of the four skills tested by the iBT (reading, listening, writing, speaking) are consistently better indicators of potential for success in language-intensive coursework than others. It will be particularly interesting to note the strength of the relationship between grades and the two sections of the IBT where the production of language is tested: writing and speaking. Strong relationships between writing and speaking scores and academic performance, especially if these relationships are stronger than those for reading and listening, would suggest that scores on proficiency tests serve as stronger

evidence of international students' academic potential when they reflect communication skills as well as comprehension and knowledge.

Participants and Data

Existing ISU records are analyzed in this study; no new data have been collected and there was no active participation from any individuals. "Participants" mentioned in this study are anonymous international students who have attended Iowa State University from the Fall 2009 term through the Spring 2011 term. Academic records and information on test scores for participants were obtained from both the Department of English and the Office of the Registrar at ISU.

Data Obtained from Department of English

The following information was requested (to the extent that it was available) from ISU's Department of English for all international students admitted to ISU during the Fall 2009, Spring 2010, Summer 2010, Fall 2010, and Spring 2011 terms:

1. Term of admission (the term that the EPT was taken). These data do not serve as a variable in any correlations, but is useful for running the same correlations for data from different terms to test the consistency with which relationships between variables occur from one term to the next.
2. Scores on the listening and speaking sections of the EPT.
3. Outcome of EPT writing test, indicating either a passing mark or placement in an ESL composition class.

In order to use EPT writing test outcomes as a variable in correlations, it was necessary to convert outcomes into numerical values. Doing so enables descriptive statistics to be calculated for the range of values indicating EPT writing performance. The values assigned to class placement range from 2 (highest) to 0 (lowest) according to the following scale:

2 : Passing mark on EPT writing / placement in English 150 if undergraduate.

1 : Placement in English 101C if undergraduate or 101D if graduate.

0 : Placement in English 101B.

Data Obtained from Office of the Registrar

Information requested of the same international students from ISU's Office of the Registrar includes:

1. Composite iBT score. This information was provided to the Office of the Registrar by the Department of Admissions.
2. Scores on individual sections of the iBT (listening, reading, speaking, and writing sections).
3. Grades obtained in the English classes that international students were placed in based on EPT writing performance. Only grades earned in classes that were taken during the term of admission were eligible for inclusion in the correlations involving grades in this study. This increases the likelihood that participants' observed performance in English classes is reflective of the same level of proficiency that is indicated by their iBT scores at the time of their admission. Information will not be available pertaining to sections and instructors of composition classes.

As with EPT writing test outcomes, letter grades in English classes are converted into numerical values. Values are assigned to grades based on Iowa State University's standard grade point scale used for calculation grade point averages:

A	: 4.00	B	: 3.00	C	: 2.00	D	: 1.00
A-	: 3.67	B-	: 2.67	C-	: 1.67	D-	: 0.67
B+	: 3.33	C+	: 2.33	D+	: 1.33	F	: 0.00

A large number of different sections of the English composition classes involved in this study are offered each term. As a result, these classes are taught by many different instructors whose standards likely vary from one individual to the next. The Department of English provides guidelines for evaluation to the instructors of these classes, but the grades on

assignments in these classes that ultimately determine each student's grade for the term depend a great deal on each individual teacher's judgment. It is therefore not possible to ensure that all of the grades noted for each English composition class reflect exactly the same evaluation standards, which is an acknowledged limitation of this study.

Office for Responsible Research Compliance

Academic records are confidential information, therefore it was necessary to obtain the approval of the IRB (Institutional Review Board) at ISU's Office for Responsible Research before the collection of data could proceed. The IRB agreed to exempt the study on the basis that there was no risk to human subjects provided that all personally identifiable information had been removed records by the time they were obtained. Student identification numbers were referenced by the Department of English and the Office of the Registrar order to match data from both sources with the same individuals. After merging the data, the Office of the Registrar then replaced all student ID numbers were then replaced with non-identifiable study ID numbers before the data were released for analysis. Another acknowledged limitation is that this study therefore must proceed under the assumption that EPT results obtained from the English department were accurately matched with grades and TOEFL scores for the same individuals.

Procedure for Statistical Analysis of Data

The analyses of data in order to answer the research questions in this study rely primarily on correlation formulae to obtain correlation coefficients. Correlation is the most widely used method of calculating the nature and strength of statistical relationships between variables (Chen & Popovich, 2002). In language testing, correlation formulae are most commonly used to investigate the relationships between scores on different tests that presumably measure some of the same abilities, though they have also been used to determine the relationships between test scores and GPA or other kinds of academic performance in studies such as those cited in Simner (1998) and Graham (1987).

Outline of Correlations in Study

Pertaining to research question #1, which concerns the relationships between iBT scores and EPT performance, the following pairs of variables are correlated:

1. Scores on the reading section of the iBT and scores on the EPT reading test.
2. Scores on the listening section of the iBT and scores on the EPT listening test.
3. Scores on the writing section of the iBT and level of English composition class placement based on EPT writing performance.
4. iBT composite scores and level of English composition class placement based on EPT writing performance.

Coefficients for the correlation of listening and reading scores on both the iBT and EPT will be calculated using the Pearson product-moment correlation coefficient. Scatterplot graphs will also illustrate the relationship between these variables. The Spearman rank-order correlation coefficient will be used to calculate coefficients in the correlations of EPT class placement and iBT writing and composite scores. All of these correlations will be run only aggregately for all five terms because most of the records containing iBT section scores come only from the Fall 2010 and Spring 2011 terms.

The Spearman formula will also be used to correlate following pairs of variables pertaining to research questions #2, which concerns the relationship between iBT scores and grades in English composition classes:

1. iBT composite scores and letter grades obtained English 150.
2. iBT composite scores and letter grades obtained English 101B.
3. iBT composite scores and letter grades obtained English 101C.
4. iBT composite scores and letter grades obtained English 101D.
5. iBT composite scores and letter grades obtained in all four English composition classes combined.
6. iBT composite scores and letter grades obtained in the three undergraduate English composition classes combined (English 150, 101B, and 101C).

Since a substantial number of records containing the necessary information for these correlations is available from each term, it will be possible to run the above correlations for

each term individually (Fall 2009, Spring 2010, Summer 2010, Fall 2010, and Spring 2011) as well as aggregately for all terms.

To obtain answers for research question #3, which addresses the relationship between scores on different sections of the iBT and grades in English composition classes, the following pairs of variables will be correlated using the Spearman formula:

1. iBT reading scores and letter grades in English composition classes.
2. iBT listening scores and letter grades in English composition classes.
3. iBT writing scores and letter grades in English composition classes.
4. iBT speaking scores and letter grades in English composition classes.

As with research question #1, that almost all of the records containing iBT section scores come from the Fall 2010 and Spring 2011 terms means that it would not be possible to run meaningful correlations for the other terms. These correlations will therefore only be run aggregately.

Descriptive Statistics

For each variable used in a correlation in this study, the following descriptive statistics will be calculated:

1. Count (N). The number of values in the distribution that is used in the correlation.
2. Mean. The average value in the distribution. When the variable is a grade, the mean will represent the GPA of the group in that particular English composition class.
3. Median. The value that represents the middle point in the distribution, where the number of values that are equal or higher is the same as the number that are equal or lower.
4. Standard Deviation. The degree of variation between values. In a "normal" distribution, most values fall within one standard deviation of the mean score, while very few values are more than two standard deviations higher or lower than the mean score.
5. Kurtosis. The peakedness of the distribution curve. Higher (positive) kurtosis values indicate that values are not distributed as evenly as in a perfectly normal distribution (a larger number of variables fall into a more limited range of values.) Low (negative)

kurtosis values indicate that the scores are more evenly distributed than is indicative of a perfectly normal distribution (a lower proportion of values fall within one standard deviation of the mean, while higher and lower values are more common.)

6. Skewness. The symmetry of the distribution curve. Negative skewness values indicate that values higher than the mean outnumber values lower than the mean, while positive skewness indicates a preponderance of lower values.

For every correlation of two variables in the study, the following statistics will be calculated:

1. Count (N). The number of pairs of values in the correlation.
2. Correlation coefficient (r/r_s). The figure most central to this study, which determines the strength and direction of the relationships between variables. The coefficient will be calculated using either the Pearson product-moment correlation (r) for comparisons of test scores. For correlations of test scores and ordinal data such as class placement or letter grades, the Spearman rank-order correlation coefficient (r_s) will be used.
3. Overlap. The correlation coefficient (whether obtained using the Pearson or Spearman formula) squared and converted to a percentage to ascertain the extent to which one variable depends on another or both variables are determined by the same underlying factors.
4. Probability (*p*). The likelihood that there is no relationship between variables and that any observed relationship therefore occurred by chance.

Descriptive statistics for distributions of variables in thus study will be calculated using JMP 9.0 and Microsoft Excel. Statistics for Pearson correlations in this study will also be calculated using JMP 9.0. For Spearman correlations in this study, statistics will be calculated using Patrick Wessa's online Spearman rank order calculator at www.wessa.net.rankcorr.wasp (2011).

Description of Analyses and Rationale

The roles of correlation formulae, overlap, statistical significance and scatterplots have already been mentioned when the procedure for the analysis of data was outlined. This section of the chapter will describe these analyses in more detail and clarify their role in the study.

Correlation Formulae

The primary means of testing the strength of the relationship between each pair of variables correlated in this study is the use of a correlation formula to obtain a correlation coefficient – a numerical value that characterizes a statistical relationship. The coefficients obtained using the two correlation formulae used in this study, the Pearson product-moment correlation coefficient and the Spearman rank-order correlation coefficient (also known as "Spearman's rho"), are interpreted similarly. Bachman (2004) explains in further detail:

> The values of both the Spearman and the Pearson correlation coefficients can range between negative one (-1.00) and positive one (+1.00), and I recommend reporting to at least three decimal points. Positive coefficients indicate direct relationships, while negative coefficients indicate inverse relationships. The larger the coefficient, positive or negative, the stronger the relationship, so that a coefficient that is close to one, either positive or negative, indicates a very strong relationship, while coefficients that are near zero indicate very weak relationships (p. 89).

The Pearson correlation formula, which is the most commonly used correlation formula in the social sciences (Bachman, 2004), will be used in this study for the correlation of reading and listening scores on the iBT and EPT in order to answer research question #1. Reading and listening scores on both tests are variables measured on an "interval" scale in that they are objectively quantifiable and original values have not been converted to another scale (Bachman, 2004). Bachman (2004) recommend the Pearson formula for the correlations when both variables represent interval data.

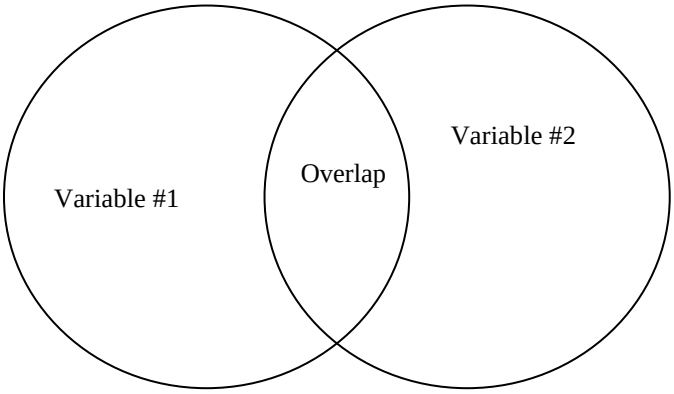
When iBT scores are correlated with class placement or letter grades, however, interval data are being correlated with "ordinal" data, where ranges of values are rated according to a scale (letter grades) or values represent ratings based on judgment (class placement). Bachman (2004) recommends the use of the Spearman correlation formula rather than the Pearson formula when interval data are correlated with ordinal data. Therefore, the vast majority of the correlation coefficients in this study will be obtained using the Spearman correlation formula (the correlation of iBT scores and class placement for research question #1 and all of the correlations of iBT scores and grades for research questions #2 and #3).

Another situation in which the Spearman formula is recommended over the Pearson formula is if values are not distributed normally for either or both of the variables in the correlation (Bachman, 2004). Therefore, the skewness and kurtosis of the distributions of values for reading and listening scores will be examined to assess their normality before these variables are correlated using the Pearson formula. If there is strong evidence of a non-normal distribution, the Spearman formula may be used for all of the correlations for research question #1 as well.

Calculation of Overlap Between Variables

A way of further analyzing the relationship between two variables that will be applied to all of the correlations in this study is to square the correlation coefficient obtained using either the Pearson or Spearman correlation formulae. The r^2 (or r_s^2 when the Spearman formula is used) is called the "coefficient of determination" (Bachman, 2004) and indicates the amount of "overlap" between two variables when converted into a percentage (Douglas, 2010). When comparing two language tests, the percentage of overlap represents the extent to which scores on each test are determined by the same underlying ability (Douglas, 2010). The concept of overlap can be demonstrated with a Venn diagram like Figure 1.

Figure 1. Overlap in language testing (adapted from Douglas, 2010)



Calculation of overlap provides additional insight into the practical significance of correlatins based on their strength (Graham, 1987). For example, a correlation coefficient of 0.600 may seem to indicate a very strong relationship compared to a coefficient of 0.200 or 0.100, but the 36% overlap that is derived from a coefficient of 0.600 means that still 64% of what determines one variable is unrelated to what determines the other variable. The calculation of overlap will therefore be a useful additional step in the analysis of these statistical relationships.

Statistical Significance

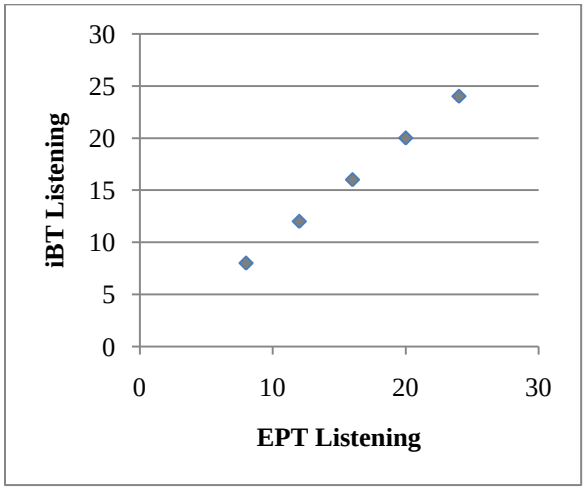
Researchers in language assessment normally report the statistical significance of their findings as well as the findings themselves (Douglas, 2010). In order to be considered statistically significant, the probability associated with a calculation should ideally be 0.01 or less. Failing that, a probability of 0.05 is still acceptable (Douglas, 2010). Calculating probability is a complex operation, although most statistics software used to calculate correlation coefficients will calculate probability as well. Generally speaking, as the strength of the relationship and sample size increase, probability decreases and the finding is more likely to demonstrate itself to be statistically significant. The desired probability of less than 0.01 indicates a less than 1% likelihood that an observed positive or negative relationship between variables occurred by chance. According to Douglas (2010), “A 5% chance of a result happening by chance is as much risk as statisticians are willing to take, and they prefer the odds to be one in one hundred. Thus, 0.05 and 0.01 are the conventionally accepted

standards for statistical significance around the world" (p. 98). As with overlap, probability will be calculated and noted for all correlations pertaining to all of the research questions in this study.

Scatterplots

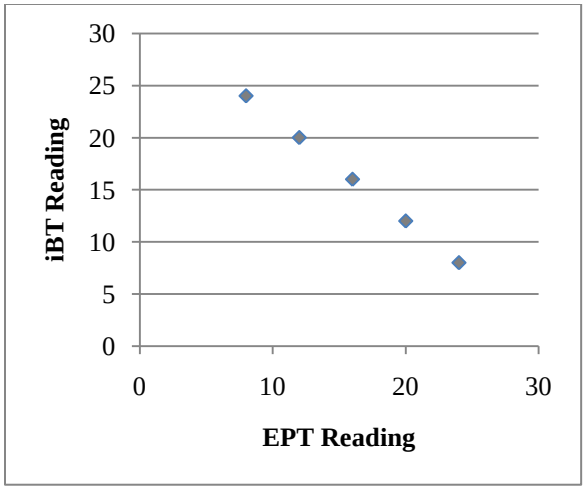
"Scatterplot" graphs, in which values for one variable are charted to the x-axis while values for the other variable are charted to the y-axis, visually represent statistical relationships. In the case of a perfectly direct (or perfectly "positive") relationship, in which high scores on one test always corresponded to equally high scores on another test, markers representing pairs of values will form a straight line from the bottom left of the graph to the upper right, as is shown in Figure 2.

Figure 2. Example scatterplot depicting direct relationship



In the opposite situation, where higher values on the x-axis correspond to proportionately lower scores on the y-axis and vice versa, there is a perfectly inverse (or perfectly "negative") relationship and markers for each pair or scores would form a straight line from the top left to the bottom right, as is shown in Figure 3. Of course, strongly negative relationships are usually not expected to actually occur in correlations of language test scores unless there is some known reason why mastering the skills that contribute to higher scores on one test would detrimentally impact one's scores on the other test.

Figure 3. Example scatterplot depicting inverse relationship



When relationships are not perfectly linear (as is normally the case), scatterplots are useful as a means of further insight into the reasons why a relationship is not perfectly linear beyond what can be inferred from a correlation coefficient alone. This is of particular relevance to the correlation of reading and listening scores on the iBT and EPT, where scatterplots will reveal if there is a pattern where incoming international students with relatively high iBT scores nevertheless score relatively low on the EPT or vice versa. A large number of cases where participants score low on the iBT but high on the EPT would merely serve as further evidence that the two tests are not measuring exactly the same skills. However, a preponderance of cases where participants have high scores on the iBT but low scores on the EPT would suggest that the EPT is doing its job by discriminating better between participants' ability levels among international students who have (in the vast majority of cases in this study) already scored 71 or higher on the TOEFL. Analysis of scatterplots will therefore contribute to the discussion of answers to research question #1.

Summary of Analyses

Now that the procedure for obtaining answers to the research questions in this study have been identified and the analyses that will be used have been explained in further detail,

the following Tables 2, 3, and 4 summarize all of the correlations that were run in this study, for which results will be discussed in Chapter 4. For every Pearson or Spearman correlation that was run, overlap and statistical significance were calculated as well.

Table 2: Summary of analyses for research question #1 – How strong is the statistical relationship between the iBT scores of incoming international students and the same individuals' performance on ISU's EPT?

Variable #1	Variable #2	Analysis	Sample	Displayed in:
iBT reading scores	EPT reading scores	Pearson Correlation	aggregate	Table 7
iBT listening scores	EPT listening scores	Pearson Correlation	aggregate	Table 7
iBT reading scores	EPT reading scores	Scatterplot	aggregate	Figure 3
iBT listening scores	EPT listening scores	Scatterplot	aggregate	Figure 4
iBT writing scores	EPT writing class placement	Spearman Correlation	aggregate	Table 8
iBT composite scores	EPT writing class placement	Spearman Correation	aggregate	Table 8

Table 3: Summary of analyses for research question #2 – How strong is the statistical relationship between international students' iBT composite scores and the grades they obtain in the English composition classes into which they are placed?

Variable #1	Variable #2	Analysis	Sample	Displayed in:
iBT composite scores	grades: English 150	Spearman Correlation	aggregate and by term	Table 12
iBT composite scores	grades: English 101B	Spearman Correlation	aggregate and by term	Table 13
iBT composite scores	grades: English 101C	Spearman Correlation	aggregate and by term	Table 13
iBT composite scores	grades: English 101D	Spearman Correlation	aggregate and by term	Table 14
iBT composite scores	grades: all English composition classes	Spearman Correlation	aggregate and by term	Table 15
iBT composite scores	grades: all undergraduate English comp. classes	Spearman Correlation	aggregate and by term	Table 16

Table 4: Summary of analyses for research question #3 – How strong is the statistical relationship between international students' iBT composite scores and the grades they obtain in the English composition classes into which they are placed?

Variable #1	Variable #2	Analysis	Sample	Displayed in:
iBT reading scores	grades: English 150	Spearman Correlation	aggregate	Table 19
iBT reading scores	grades: English 101B	Spearman Correlation	aggregate	Table 19
iBT reading scores	grades English 101C	Spearman Correlation	aggregate	Table 19
iBT reading scores	grades: English 101D	Spearman Correlation	aggregate	Table 19
iBT reading scores	grades: all undergraduate English comp. classes	Spearman Correlation	aggregate	Table 19
iBT reading scores	grades: all English composition classes	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: English 150	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: English 101B	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: English 101C	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: English 101D	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: all undergraduate English comp. classes	Spearman Correlation	aggregate	Table 19
iBT listening scores	grades: all English composition classes	Spearman Correlation	aggregate	Table 19
iBT writing scores	grades: English 150	Spearman Correlation	aggregate	Table 20
iBT writing scores	grades English 101B	Spearman Correlation	aggregate	Table 20
iBT writing scores	grades: English 101C	Spearman Correlation	aggregate	Table 20
iBT writing scores	grades: English 101D	Spearman Correlation	aggregate	Table 20
iBT writing scores	grades: all undergraduate English comp. classes	Spearman Correlation	aggregate	Table 20
iBT writing scores	grades: all English composition classes	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: English 150	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: English 101B	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: English 101C	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: English 101D	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: all undergraduate English comp. classes	Spearman Correlation	aggregate	Table 20
iBT speaking scores	grades: all English composition classes	Spearman Correlation	aggregate	Table 20

Issues in the Interpretation of Findings in Correlation Studies

Some degree of caution is always necessary when analyzing correlation data, as there is a wide variety of potential issues that can lead to "biased" and therefore misleading results during correlation studies. One issue affecting all of the correlations pertaining to all of the research questions in this study is the restricted range of iBT scores that will be correlated with other variables due to ISU's admissions policy. Because this study involves ISU students, and a minimum composite iBT score of 71 is required for admission to ISU, iBT scores in the data used in this study will always be 71 or higher except in rare cases where individuals who had lower iBT scores were nevertheless admitted according to other criteria. A small number of iBT scores below 71 will therefore be included in the study, but the vast majority of scores correlated in this study will fall into the low-intermediate to high range of 71 or higher. Participants in this study therefore represent somewhat of a "truncated sample" in terms of their iBT composite scores. The range of scores on sections of the iBT will be impacted as well, because individuals who meet the minimum composite score required for admission are also less likely to have very low scores on one or more sections of the iBT. Correlations involving truncated samples usually produce weaker coefficients than would have been observed for a wider range of values (Bachman, 2004).

Possibly the greatest issue pertaining to the theoretical significance of correlation studies is that a direct statistical relationship between two variables, even when it demonstrates itself to be statistically significant, does not necessarily prove the existence of a causal relationship between the variables. This is because statistical relationships are in no way indicative of the extent to which one variable is determined by another (Hays, 1994, cited in Chen and Popovich, 2002). Although a strong direct relationship is good evidence to support an already existing hypothesis that a causal relationship between two variables does exist, it is never definitive proof (Chen and Popovich, 2002). In this study, a causal relationship is believed to exist between the English proficiency that is measured by the TOEFL (albeit not the TOEFL scores themselves) and performance on the EPT and in classes. However, even if very strong relationships and significant amounts of overlap are

observed in the results, they will not represent actual proof that iBT scores and other variables coincide because of the role of English proficiency and not for other reasons.

CHAPTER 4.
RESULTS AND ANALYSIS

The dominance of the TOEFL iBT in English proficiency testing, at least in the case of students planning to study in the United States, is attested to by the high percentage of iBT scores relative to scores on other proficiency tests among the 1432 international students whose records were obtained from ISU's Office of the Registrar (as shown in Table 5 below). More than half of the international students admitted to Iowa State University from Fall 2009 through Spring 2011 reported an iBT score, though a sizeable percentage also reported an IELTS score, reflecting the growing acceptance among universities and usage of this test. The significant number of PBT scores mostly reflects international students who took the TOEFL at ISU. Participants in ISU's Intensive English and Orientation Program are given a chance to take the TOEFL and qualify for admission at the end of each term. International applicants whose proficiency test scores are very close to but do not quite meet the admission requirement are also given a chance to take the TOEFL and meet the admission requirement upon their arrival at ISU. In these cases, where the TOEFL is administered directly by an institution, the PBT must be administered rather than the iBT because the iBT can only be taken at ETS test centers. There were also many records for which both iBT and IELTS scores were reported. Exact counts are displayed in Table 5. Because some of the participants in this study reported scores on more than one test, the summation of the different test types exceeds the number of student records analyzed. Of the total number of student records analyzed, 108 had no test score reported

Table 5. Test scores reported for international students admitted Fall 2009 – Spring 2011.

Test Type	Number	% of Total
TOEFL PBT	246	17.18%
TOEFL iBT	839	58.59%
IELTS	338	23.60%
No valid test	108	7.54%
Total Records	1432	100%

Of the above 839 records containing an iBT score, the number that will qualify for inclusion in the analyses to answer each research question will depend on how many of those records contain the necessary additional information on other variables. In the following sections of this chapter, the results for each research question in this study will then be displayed, discussed, and summarized. Descriptive statistics for the ranges of variables used in each correlation are reported along with the calculated results of analyses of the statistical relationships.

Research Question #1

How strong is the statistical relationship between the iBT scores of incoming international students and the same individuals' performance on ISU's EPT?

In order to be included in the correlations related to research question #1, which compares performance on the iBT and corresponding sections of the EPT, it was necessary that records containing an iBT score also include the breakdown of the composite score into the four different sections of the iBT: reading, listening, writing and speaking. Of the 839 records containing an iBT score, 338 (overwhelmingly from the Fall 2010 and Spring 2011 terms) contained this breakdown of scores. Correlations involving iBT section scores are run for all five terms not by individual semester of admission but aggregately because there would not be a sufficiently count of participants for the Fall 2009, Spring 2010, and Summer 2010 terms to make any meaningful inferences from the correlation coefficients obtained. The count for the Spearman correlations of iBT writing and composite scores with class placement based on the EPT writing test is 324 because 16 of the aforementioned 338 records that lacked the necessary information pertaining to class placement.

Descriptive Statistics

Descriptive statistics for each range of variables used in the analysis of the relationship between iBT scores and EPT performance are displayed in Table 6. Both the iBT and EPT reading and listening test sections contain 30 items.

Table 6. Descriptive statistics for variables used in iBT/EPT correlations

Variable	N	Mean	Median	Standard Deviation	Kurtosis	Skewness
iBT Reading	338	22.51	23	4.89	0.24	-0.69
iBT Listening	338	20.93	21	4.71	-0.01	-0.44
iBT Writing	324	21.56	21	3.49	0.19	-0.35
iBT Composite	324	84.77	84	11.16	0.49	-0.40
EPT Reading	338	17.27	18	4.85	-0.39	-0.32
EPT Listening	338	16.04	16	4.48	-0.43	-0.11
EPT Writing (Class Placement)	324	1.20	1.00	0.64	-0.64	-0.20

The median scores for iBT reading and listening (23/30 and 21/30, respectively) are noticeably higher than those for EPT reading and listening (18/30 and 16/30, respectively). This is likely a result of the restricted range of iBT scores that international students admitted to ISU represent. Because a composite iBT score of 71 is required for admission to ISU, and it is less likely that examinees with low scores on any section of the iBT will attain a composite score of 71, international students who have been admitted to ISU are unlikely to report very low scores on any section of the iBT. Meanwhile, the lower mean and median scores for EPT reading and listening reflect a higher incidence of low scores (relative to the maximum score possible) on these EPT sections than on the corresponding iBT sections. This suggests that the EPT has been successful at more precisely discriminating between individual ability levels within the specific proficiency range of international students who have been admitted to ISU (whose iBT scores are 71 or higher).

The skewness and kurtosis of the distributions of scores on iBT reading, iBT listening, EPT reading, and EPT listening reveal that the distribution of values for each is sufficiently normal that the Pearson formula may be used as planned to correlate these variables. According to Bachman (2004), skewness and kurtosis values ranging from -2.00 to +2.00 are indicative of a "reasonably normal" distribution "as a rule of thumb" (p. 74). The distributions for iBT reading and listening are more negatively skewed than those of

corresponding EPT sections due to a greater frequency of very high scores within the iBT distributions. Nevertheless, skewness and kurtosis remain well within the desired range of -2.00 to +2.00 for all four variables that will figure into Pearson correlations.

Correlations of iBT and EPT Reading and Listening Scores

Results for the correlation of reading and listening scores on the EPT and iBT are displayed in Table 7 below.

Table 7. Pearson correlations of iBT and EPT scores – listening and reading

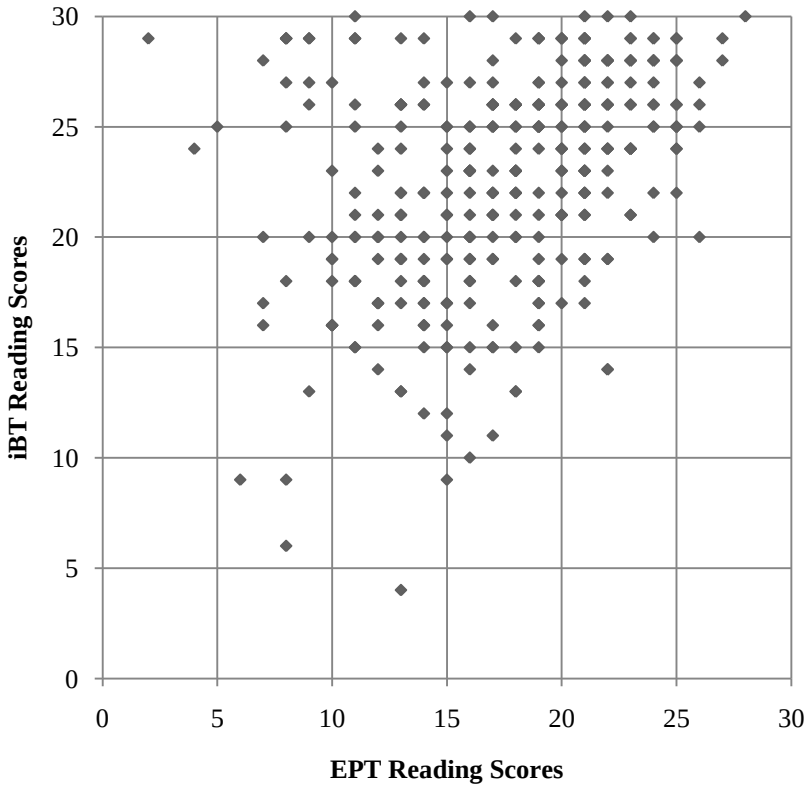
Variable #1	Variable #2	N	r	Overlap	*p*
iBT Reading Score	EPT Reading	338	0.363	13.18%	< 0.01
iBT Listening Score	EPT Listening	338	0.413	17.06%	< 0.01

The coefficients of 0.363 (reading) and 0.413 (listening) obtained for the correlation of iBT and EPT scores indicate that a relationship of moderate strength exists between these sections of the two tests. However, the relationship is not nearly strong enough to infer from these results that both tests are measuring the same specific skills. The overlap between the iBT and EPT is 13.18% for the reading sections and 17.06% for the listening sections, indicating that the vast majority of an individual's performance on the reading and listening sections of the EPT owes to factors other than the level of English proficiency that is reflected by iBT scores pertaining to similar skills. However, a stronger relationship may have been observed had there been more varied range of iBT scores in the distributions of scores used in these correlations.

Scatterplots of iBT and EPT Reading and Listening Scores

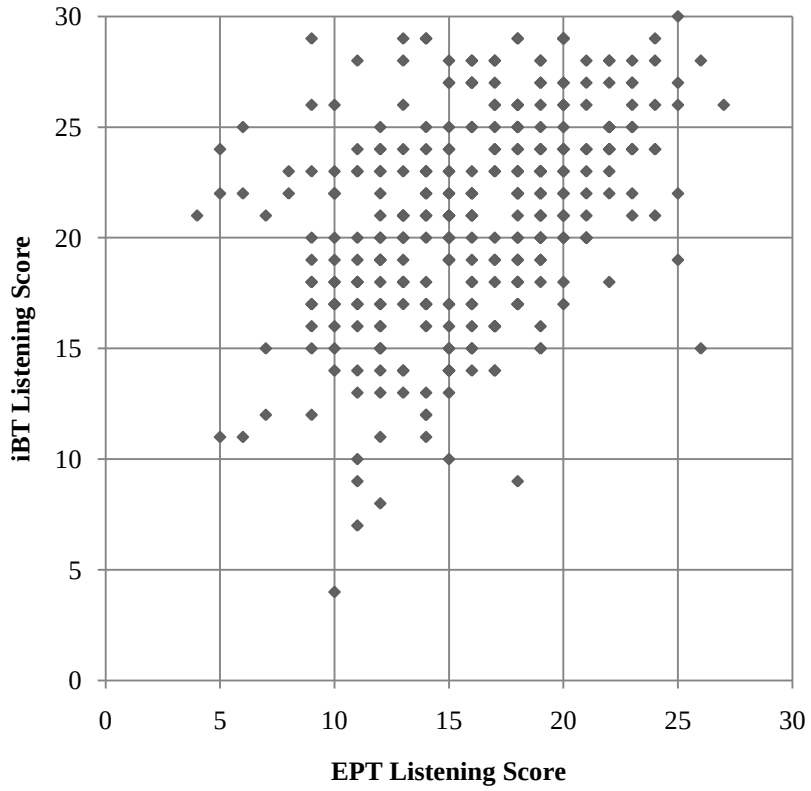
Scores on the reading section of the iBT and the EPT reading test in the data used in this study are charted in Figure 4 below. EPT reading scores are charted to the x-axis while iBT reading scores are charted to the y-axis.

Figure 4. Scatterplot of iBT reading scores and EPT reading scores



The strong clustering of markers in the upper right quadrant of Figure 4 is indicative of the direct relationship reflected in the correlation coefficient of 0.363. There is a certain tendency for high iBT reading scores to coincide with high EPT reading scores, but the presence of many markers in the upper left quadrant of the plot area indicates that a large number of participants with relatively high scores on iBT reading nevertheless obtained relatively low scores on EPT reading. A participant's iBT reading score therefore cannot be considered an entirely accurate indicator of how well the same participant will fare EPT reading test. A similar pattern can be noted in Figure 5, a scatterplot with scores on the EPT listening test on the x-axis and scores on the iBT listening section on the y-axis.

Figure 5. Scatterplot of iBT listening scores and EPT listening scores



The slightly stronger relationship between listening scores (r=0.413) is reflected in less dense clustering of markers in the upper left quadrant of Figure 5, though once again the vast majority of the markers are in the upper left or upper right of the plot area. The bottom right quadrant is relatively empty, showing that participants who scored high on the EPT but low on the iBT were uncommon compared to the other way around, as was also the case when reading sections were compared. One possible explanation for the frequent disparity in performances on both tests is that it is not possible for incoming international students to "cram" for the EPT the same way they are likely to have done before taking the iBT. While TOEFL preparation materials are plentiful and TOEFL preparation may be a major goal in EFL instruction (Hamp-Lyons, 1998), international students will know relatively little about the content of the EPT before taking it upon their arrival.

It is evident from both the correlation coefficients observed and the examination of the scatterplots that an incoming international student's reading and listening scores on the iBT cannot be depended on to accurately forecast how well the same individual will perform

on the EPT reading and listening tests. One could nevertheless claim that a high iBT reading or listening score is evidence that the same individual is more likely to obtain a high EPT reading or listening score.

Correlations of iBT Writing and Composite Scores and Class Placement

Table 8 below displays the results of the correlations of iBT writing and composite scores and class placement based on the rating of essays on the EPT writing test.

Table 8. Spearman correlations of iBT scores and EPT writing performance

Variable #1	Variable #2	N	r_s	Overlap	*p*
iBT Writing Score	EPT Writing (Class Placement)	324	0.317	11.42%	0.02
iBT Composite Score	EPT Writing (Class Placement)	324	0.406	14.90%	< 0.01

The correlation coefficient for iBT writing scores and class placement ($r_s = 0.354$) is sufficient to demonstrate that a positive relationship exists between these variables but the relationship is not tellingly strong. The overlap of 11.42% between these variables indicates that nearly 90% of the factors that determine performance on the EPT writing test operate independently of the skills measured by the writing section of the iBT. What is interesting (though counter-intuitive) is the slightly stronger relationship ($r_s = 0.395$) between composite iBT scores and class placement. EPT writing tasks are not designed to be integrative the same way some tasks on the writing section of the iBT are, in which reading and listening skills also play a prominent role in the writing task. It is possible, however, that comprehension of spoken instructions and writing prompts play a role in EPT writing performance, especially in the case of graduate students where the writing prompt concerns the interpretation of visual information such as charts.

Summary

Positive, statistically significant relationships have been revealed to exist between performance on ISU's EPT and scores on sections of the iBT that test similar skills.

However, the strength of the relationships is insufficient to demonstrate that both the iBT and EPT are testing the same skills to such an extent that iBT scores could be dependably used in the same role as EPT scores. High iBT reading and listening scores, since they often coincide with low EPT reading and listening scores, cannot be relied on to predict which students will need to be placed into English 99R or 99L. Whatever the reason for the stronger relationship that has been observed between composite scores and class placement, this finding serves as some justification for the current policy of exempting international students from the EPT on the basis of particularly high composite iBT scores (instead of based on iBT writing scores specifically.) On the other hand, the fair strength of this relationship means that even a very high composite iBT score of 105 or more cannot be considered proof that the same individual would pass the writing test were he or she required to take the EPT. While drawing any conclusions or formulating any hypothesis based on these findings, one must bear in mind how the limited range of intermediate to high iBT scores is likely to have had a weakening effect on these relationships.

Research Question #2

How strong is the statistical relationship between international students' iBT composite scores and the grades they obtain in the English composition classes into which they are placed?

Of the 839 international students for whom an iBT score was noted, 503 met the criteria necessary for inclusion in the correlations of composite TOEFL scores and grades obtained in English composition classes. Most of the records that had to be excluded from these correlations were those of international students who were not enrolled in the English composition class of their placement until a later semester. A handful of other records had to be excluded when no letter grade was obtained in the English composition class in question. These include cases where the class was taken on a pass/no-pass basis, the student was given an "incomplete", or no grade was reported. International graduate students who were exempted from the EPT or whose performance on the EPT writing test was sufficient to

"pass" were not required to take an English composition class and also do not figure into these correlations.

Descriptive Statistics

Table 9 compares the GPA of international students in their English composition classes, both aggregately and grouped according to semester of admission.

Table 9. English composition GPA and mean iBT composite scores by term

Term	N	GPA	Mean iBT Score
Fall 2009	187	3.43	83.20
Spring 2010	47	3.31	80.99
Summer 2010	29	3.14	82.86
Fall 2010	196	3.17	82.31
Spring 2011	44	3.39	81.75
Aggregate	503	3.30	82.43

The figures in Table 9 reveal an apparent relationship between TOEFL scores and grades obtained in classes even before the relationship between these variables is tested with the Spearman formula. The group with the highest mean iBT score, Fall 2009, also exhibited the highest GPA in English composition classes taken during the semester of admission. However, the group with the lowest mean GPA, Summer 2010, did not also report the lowest mean iBT score. In fact, the mean iBT score for Summer 2010 is the second highest among the groups noted above. This information indicates that some logical relationship between success on English proficiency tests and success in English classes, but that this relationship is far from perfectly linear.

A complete list of the descriptive statistics pertaining to both iBT composite scores and grades sorted by English class of placement and semester of admission can be found in Tables 10 and 11 on the following two pages. Included in these tables are statistics for groups of classes, whether all English composition classes in this study (English 150, 101B, 101C, and 101D) or the three undergraduate classes (English 150, 101B, and 101C – though in rare

cases some graduate students were also placed in English 101B). Aggregate statistics (for all five terms combined) for each class and group of classes are listed as well.

Table 10. Descriptive statistics for distributions of composite iBT scores

English Class(es) of Placement	Term	N	Mean	Median	Standard Deviation	Kurtosis	Skewness
150	Fall 2009	21	85.10	83	10.89	-1.15	-0.04
150	Spring 2010	3	87.67	93	11.93	N/A	-1.61
150	Summer 2010	22	82.50	79.5	9.11	0.147	0.91
150	Fall 2010	28	89.04	87.5	10.50	-1.39	-0.14
150	Spring 2011	4	90.5	91.5	3.87	2.36	-1.38
150	**Aggregate**	**78**	**86.15**	**85.5**	**10.23**	**-1.17**	**0.09**
101B	Fall 2009	31	73.61	76	14.71	1.49	-0.95
101B	Spring 2010	3	79.33	80	10.01	N/A	-0.30
101B	Summer 2010	0	N/A	N/A	N/A	N/A	N/A
101B	Fall 2010	28	78.46	77	13.34	0.76	-0.07
101B	Spring 2011	13	85.46	85	7.93	-0.93	0.02
101B	**Aggregate**	**75**	**77.71**	**78**	**13.54**	**1.60**	**-0.76**
101C	Fall 2009	59	76.27	76	10.68	1.98	-0.83
101C	Spring 2010	34	78.50	78	6.46	0.38	0.91
101C	Summer 2010	3	79.00	73	12.17	N/A	1.68
101C	Fall 2010	92	78.39	78.5	9.97	3.87	-0.90
101C	Spring 2011	21	75.29	76	9.64	2.53	-0.97
101C	**Aggregate**	**209**	**77.51**	**77**	**9.68**	**2.99**	**-0.80**
101D	Fall 2009	76	91.97	92	7.59	0.44	-0.61
101D	Spring 2010	7	86.14	85	8.84	1.76	0.60
101D	Summer 2010	4	87.75	87.5	4.92	1.35	0.30
101D	Fall 2010	48	88.13	87	6.32	-1.11	0.18
101D	Spring 2011	6	90.50	91	8.48	-1.63	-0.23
101D	**Aggregate**	**141**	**90.19**	**91**	**7.41**	**-0.38**	**-0.20**
All Undergraduate	Fall 2009	111	77.20	77	12.52	1.78	-0.78
All Undergraduate	Spring 2010	40	79.25	78.5	7.33	-0.20	0.79
All Undergraduate	Summer 2010	25	82.08	78	9.30	-0.15	0.82
All Undergraduate	Fall 2010	148	80.42	80	11.49	1.65	-0.35
All Undergraduate	Spring 2011	38	80.37	81	10.33	1.41	-0.73
All Undergraduate	**Aggregate**	**362**	**79.41**	**78**	**11.25**	**1.97**	**-0.54**
All Classes	Fall 2009	187	83.20	84	12.99	1.41	-0.88
All Classes	Spring 2010	47	80.28	79	7.86	0.03	0.77
All Classes	Summer 2010	29	82.86	82	8.98	-0.40	0.60
All Classes	Fall 2010	196	82.31	81.5	10.96	1.79	-0.57
All Classes	Spring 2011	44	81.75	82	10.61	1.19	-0.64
All Classes	**Aggregate**	**503**	**82.43**	**82**	**11.39**	**1.56**	**-0.63**

Table 11. Descriptive statistics for distributions of grades in composition classes

English Class(es) of Placement	Term	N	Mean (GPA)	Median	Standard Deviation	Kurtosis	Skewness
150	Fall 2009	21	3.05	3.00	0.89	6.27	-2.07
150	Spring 2010	3	3.56	3.67	0.51	N/A	-0.95
150	Summer 2010	22	2.91	3.00	0.43	3.17	-0.64
150	Fall 2010	28	3.17	3.33	0.67	-1.41	-0.22
150	Spring 2011	4	3.42	3.50	0.57	0.27	-0.74
150	**Aggregate**	**78**	**3.09**	**3.00**	**0.68**	**4.26**	**-1.23**
101B	Fall 2009	31	3.11	3.33	0.80	2.81	-1.60
101B	Spring 2010	3	2.67	2.67	1.34	N/A	-0.01
101B	Summer 2010	0	N/A	N/A	N/A	N/A	N/A
101B	Fall 2010	28	2.87	3.33	1.25	0.02	-1.06
101B	Spring 2011	13	3.41	3.33	0.49	-1.25	-0.09
101B	**Aggregate**	**75**	**3.05**	**3.33**	**0.98**	**1.57**	**-1.42**
101C	Fall 2009	59	3.25	3.67	0.80	4.25	-1.84
101C	Spring 2010	34	3.26	3.33	0.85	5.51	-1.99
101C	Summer 2010	3	3.67	3.67	0.34	N/A	-0.04
101C	Fall 2010	92	2.90	3.33	1.00	1.28	-1.27
101C	Spring 2011	21	3.22	3.67	0.96	5.54	-2.06
101C	**Aggregate**	**209**	**3.10**	**3.33**	**0.92**	**2.53**	**-1.57**
101D	Fall 2009	76	3.81	4.00	0.52	38.18	-5.57
101D	Spring 2010	7	3.71	4.00	0.49	-0.84	-1.23
101D	Summer 2010	4	4.00	4.00	0.00	N/A	N/A
101D	Fall 2010	48	3.84	4.00	0.36	4.22	-2.31
101D	Spring 2011	6	3.89	4.00	0.27	N/A	N/A
101D	**Aggregate**	**141**	**3.83**	**4.00**	**0.45**	**36.32**	**-5.04**
All Undergraduate	Fall 2009	111	3.17	3.33	0.82	3.82	-1.76
All Undergraduate	Spring 2010	40	3.24	3.33	0.86	4.05	-1.77
All Undergraduate	Summer 2010	25	3.00	3.00	0.48	1.79	-0.27
All Undergraduate	Fall 2010	148	2.95	3.33	1.00	1.19	-1.24
All Undergraduate	Spring 2011	38	3.31	3.50	0.78	7.52	-2.18
All Undergraduate	**Aggregate**	**362**	**3.09**	**3.33**	**0.89**	**2.56**	**-1.52**
All Classes	Fall 2009	187	3.43	3.67	0.78	6.14	-2.24
All Classes	Spring 2010	47	3.31	3.67	0.83	4.53	-1.85
All Classes	Summer 2010	29	3.14	3.00	0.57	0.26	-0.10
All Classes	Fall 2010	196	3.17	3.33	0.96	1.91	-1.47
All Classes	Spring 2011	44	3.39	3.67	0.76	8.07	-2.28
All Classes	**Aggregate**	**503**	**3.30**	**3.67**	**0.86**	**3.60**	**-1.77**

The vast majority of distributions of iBT composite scores in Table 10 are negatively skewed. This means that a preponderance of these 503 participants obtained higher iBT scores than the mean score of 82.43 for this group. The English 150 group is the only group whose aggregate distribution of scores is positively skewed, though at 0.09 only very

marginally so. It seems logical that the highest mean scores would occur among the graduate students who are placed in English 101D, as the required iBT score for graduate admission to ISU is 79 or higher depending on the program. Mean scores for undergraduates placed in English 150 group are higher than the mean scores for the English 101B and 101C groups, reflecting the previously noted positive relationship between iBT scores and class placement. What is surprising, however, is that the aggregate mean score for participants placed in English 101B is marginally higher than that of those placed in English 101C. This likely owes to the group of participants placed in English 101B in Spring 2011, whose mean iBT score of 85.46 is comparable to the aggregate mean score of 86.15 for those placed in English 150.

Distributions of grades for all groups in Table 11 are all negatively skewed to varying degrees, though the distributions for English 101D are by far the most heavily negatively skewed. That the median score for all English 101D groups is 4.00 (indicating an A letter grade), combined with the very high aggregate kurtosis of the distributions for English 101D (mostly thanks to Fall 2009), indicates that the vast majority of students who were placed in English 101D got an A. This makes sense if higher English proficiency (reflected in the high mean iBT score for the 101D group) improves the likelihood that a participant gets an A. On the other hand, the high incidence of the highest possible grades indicates that individual variance in iBT scores may not have made much of a difference. The relatively lower degrees of skewness and kurtosis for the distributions of grades for undergraduate classes may therefore indicate a better range of grades in these distributions to serve as a better basis for correlation with scores.

Correlations: English 150

Statistics for correlations of composite iBT scores and grades in English 150, the only non-ESL class correlated with iBT scores in this study, are displayed aggregately and by term in Table 12.

Table 12. Spearman correlations of iBT composite scores and grades in English 150

English Class of Placement	Term	N	r_s	Overlap	*p*
150	Fall 2009	21	0.098	0.96%	0.62
150	Spring 2010	3	1.000	100.00%	0.15
150	Summer 2010	22	0.366	13.40%	0.02
150	Fall 2010	28	0.244	5.95%	0.17
150	Spring 2011	4	0.00	0.00%	0.99
150	Aggregate	78	0.304	9.24%	< 0.01

Except in the case of the 4 participants enrolled in English 150 in Spring 2011, a positive relationship is consistently noted between iBT scores and grades in English 150. However, the combination of sample size and coefficient strength necessary for a statistically significant relationship is only present for the Summer 2010 group and when all terms are aggregated. The aggregate correlation coefficient of 0.304 for the English 150 group may not be impressive in its own right as it only represents a 9.24% overlap between the variables, but it is statistically significant at the sample size of 78 and nevertheless stronger than most of the coefficients noted in previous studies involving the PBT. When the distribution is broken up by term, however, the very small sample sizes for the Spring 2010 and Spring 2011 terms makes it impossible to obtain a statistically coefficient (the perfectly direct relationship noted for Spring 2010 and the perfectly non-existent relationship for Spring 2011 attest to the unpredictability of statistical relationships when sample sizes are so small). The relationships are also not strong enough to be considered statistically significant at the sample sizes for Fall 2009 and Fall 2010.

Correlations: English 101B and English 101C

English 101B and 101C, the two undergraduate ESL classes, constitute the majority of grades that will be correlated with iBT composite scores in this study. Statistics for correlations of composite iBT scores and grades in English 101B and 101C are displayed aggregately and by term in Table 13.

Table 13. Spearman correlations of iBT composite scores and grades in English 101B and 101C

English Class	Term	N	r_s	Overlap	*p*
101B	Fall 2009	31	0.347	12.04%	0.05
101B	Spring 2010	3	0.500	25.00%	0.48
101B	Summer 2010	0	N/A	N/A	N/A
101B	Fall 2010	28	-0.316	9.99%	0.12
101B	Spring 2011	13	0.284	8.07%	0.28
101B	**Aggregate**	**75**	**0.042**	**0.18%**	**0.63**
101C	Fall 2009	59	0.150	2.25%	0.17
101C	Spring 2010	34	0.395	15.60%	0.02
101C	Summer 2010	3	-1.000	100.00%	0.15
101C	Fall 2010	92	0.021	0.04%	075
101C	Spring 2011	21	0.009	0.01%	0.91
101C	**Aggregate**	**209**	**0.093**	**0.86%**	**< 0.01**

The aggregate correlations reveal very weak, marginal relationships between iBT composite scores and grades in both English 101B and 101C. The aggregate figure is not statistically significant for English 101B. The sample size of 209 for English 101C is sufficient for statistical significance despite the weak coefficient, however. A positive relationship between iBT scores and grades in English 101C does evidently exist, but with an overlap of less than 1% the relationship is of no practical relevance. That being noted, the also statistically significant relationship between scores and 101C grades for Spring 2010 ($r = 0.395$) indicates a 15.60% overlap between the variables. English proficiency therefore played a meaningful role in the degree of success of participants in that particular group, though still far from an overwhelmingly important role. The situation is similar for the statistically significant (at the 5% level) positive relationship noted for Fall 2009 101B grades and iBT scores, which indicates a 12.04% overlap between the variables. However, the next largest 101B group, Fall 2010, reports a negative coefficient indicating an inverse relationship between grades and iBT scores. This suggests that the relationship between iBT scores and grades in these classes can vary significantly from one sample to the next and that it is therefore unsafe to make generalizations based on these observations. Overall, with the weakness of the aggregate coefficients and the fluctuation in the strength of the relationships, from one term

to the next, it does not seem possible to forecast anything with any certainty about international students' potential for success in English 101B or 101C based on iBT scores.

Correlations: English 101D

English 101D, comprising a large portion of the total grades correlated with iBT scores in this part of the study, is somewhat of a special case due to the very high incidence of the highest possible grades in these classes. Statistics for correlations of composite iBT scores and grades in English 101D are displayed aggregately and by term in Table 14.

Table 14. Spearman correlations of iBT composite scores and grades in English 101D

Correlated Variable	Term	N	r_s	Overlap	*p*
101D	Fall 2009	76	0.153	2.34%	< 0.01
101D	Spring 2010	7	-0.080	0.64%	0.74
101D	Summer 2010	4	0.00	0.00%	0.38
101D	Fall 2010	48	-0.127	1.61%	0.27
101D	Spring 2011	6	-0.264	6.97%	0.65
101D	Aggregate	141	0.032	0.10%	< 0.01

Judging by the results of this study, there is little reason to presume that iBT scores are in any way indicative of potential performance in English 101D. This is because, as noted during the discussion of the distributions of grades, almost all participants earned an A in English 101D regardless of variance in iBT scores among individuals. The slightly negative relationship for the Spring 2010, Fall 2010, and Spring 2011 groups are evidence that the relatively few participants who did not obtain an A in English 101D did not necessarily have lower iBT scores relative to those who did get an A.

The very heavily negatively skewed distributions for Fall 2009 an all terms combined could result in a situation called heteroscedasticity, where the relationship is difficult to characterize mathematically because one distribution is heavily skewed while the other is not. This is known to have a weakening effect on correlation coefficients (Bachman, 2004). The overwhelming preponderance of As in these distributions represents an extreme case of heteroscedasticity; there simply is not enough of a range of values in one distribution to facilitate a meaningful correlation with the other range of variables.

Correlations: Combined English Composition Classes

The correlation of grades in all English composition classes together represents an "ecological" correlation in which the potential of obtaining a biased and therefore misleading coefficient must be taken into consideration. When mixed into the same correlation, multiple groups that inhabit different score ranges may create a new pattern that artificially facilitates a higher correlation coefficient than that of each group individually (Bachman 2004). For example, in the data used in this study, international graduate students placed into English 101D have higher iBT scores on average than those placed into English 150, who in turn are likely to have higher scores on average than those placed into English 101C or 101B. This could result in a trend among all four groups that is not represented in each group individually. Furthermore, the grades earned in these English classes do not take the level of the class into account. Participants who earn a B or C in English 150 may have higher iBT scores than participants who earn an A in English 101B, for example (a likely scenario since a moderately strong positive relationship between iBT scores and class placement has already been observed while answering research question #1). This will result in a greater variety of proficiency levels (and therefore iBT scores) that correspond to each letter grade, weakening the statistical relationship.

Statistics for correlations of composite iBT scores and grades in all four English composition classes correlated in this study (English 150, 101B, 101C, and 101D) are displayed aggregately and by term in Table 15.

Table 15: Spearman correlations of iBT composite scores and grades in all English composition classes combined

English Classes	Term	N	r_s	Overlap	*p*
150, 101B, 101C, 101D	Fall 2009	187	0.409	16.73%	< 0.01
150, 101B, 101C, 101D	Spring 2010	47	0.434	18.84%	< 0.01
150, 101B, 101C, 101D	Summer 2010	29	0.287	8.24%	0.05
150, 101B, 101C, 101D	Fall 2010	196	0.176	3.10%	< 0.01
150, 101B, 101C, 101D	Spring 2011	44	0.147	2.16%	0.25
150, 101B, 101C, 101D	Aggregate	503	0.291	8.47%	< 0.01

Some of the coefficients noted in Table 15 are indeed much stronger than one would expect when many of the relationships between grades and individual classes were found to be only marginal. Also, though the strength of the relationship may still vary quite a bit from one semester to the next, the relationships are far more consistent in that they are all greater than 0.1 and do not fluctuate between positive and negative coefficients. At 0.291, the aggregate coefficient for all classes, if this accurately reflects the relationship between iBT scores and English class performance, does suggest that the iBT is a better indicator of potential for academic success than was the PBT, for which correlations with grades usually produced coefficients of less than 0.200.

That the relationships for all four classes combined are often stronger than the relationships noted for individual classes for the same term suggests that the ecological correlation may have more to do with the coefficients observed than an actual link between English proficiency as measured by the TOEFL and academic performance. This is most likely the case especially because of the influence of the large group of participants in English 101D. Inclusion of the English 101D group means 141 pairs of variables are factored into the correlation for which the grade obtained is usually an A and the iBT score is relatively high (as international graduate students generally have higher iBT scores).

The pairs of values for the English 101D group do reflect a genuine situation where high grades coincide with higher iBT scores, in which case their inclusion in the correlation helps illustrate a pattern that really does exist. On the other hand, one must keep in mind how correlation coefficients are blind to causality; it is not necessarily because of their higher iBT scores that international graduate students obtain the highest possible grades with such frequency in English 101D, especially considering the weak, fluctuating relationships when scores are correlated with English 101D grades alone. The differences in the results of the separate correlation of the three undergraduate classes (English 150, 101B, and 101C) shown in Table 16 reveal the strengthening influence of the 101D group when it is included.

Correlations: Combined Undergraduate English Composition Classes

Statistics for correlations of composite iBT scores and grades in the three undergraduate English composition classes correlated in this study (English 150, 101B, 101C) are displayed aggregately and by term in Table 16.

Table 16: Spearman correlations of iBT composite scores and grades in undergraduate English composition classes.

English Classes	Term	N	r_s	Overlap	*p*
150, 101B, 101C	Fall 2009	111	0.150	2.25%	0.08
150, 101B, 101C	Spring 2010	40	0.449	20.16%	< 0.01
150, 101B, 101C	Summer 2010	25	0.119	1.42%	0.16
150, 101B, 101C	Fall 2010	148	-0.192	3.69%	0.10
150, 101B, 101C	Spring 2011	38	0.052	0.27%	0.65
150, 101B, 101C	Aggregate	362	0.081	0.66%	0.07

The exclusion of English 101D from the correlation causes the aggregate coefficient to drop precipitously from 0.291 to 0.081. The relationship is also weakened for each term except for Spring 2010, where there is a slight increase in the strength of the relationship. The greatly increased number of very high grades paired with high iBT scores when English 101D is included in the correlation does indeed facilitate a much stronger combined relationship due to a stronger clustering of values towards the higher end of both ranges of variables.

Summary

In correlations of composite iBT scores with grades obtained in English composition classes, there is considerable variance in the strength and sometimes direction of the relationship depending on how participants are grouped. At 0.291, the aggregate coefficient for all composition classes combined seems to be at least an improvement over the coefficients that were noted in many of the previous correlation studies involving the PBT. However, the pairs of high iBT scores and high grades belonging to the English 101D must be included in the correlation in order to establish even this moderate-strength relationship. When the 101D group is excluded from the correlation, the situation is similar to what was observed for most of the English composition classes individually when correlated with iBT

scores: the relationship is unpredictable and usually too weak to indicate that English proficiency, as measured by the iBT version of the TOEFL, has played a major role in deciding these international students' success or failure in these classes. This research therefore suggests that academic success, even in English composition classes, where the production of language and the integrative use of skills would theoretically play a particularly important role, is overwhelmingly determined by factors other than level of English proficiency. However, though the findings were not always statistically significant, international students' performance in English 150, the only non-ESL class of the four English composition classes used in these correlations, did correlate positively with iBT scores with greater consistency than was noted for other individual classes.

Research Question #3

How strong is the statistical relationship between international students' scores on the four different sections of the iBT (reading, listening, speaking, and writing) and the grades they obtain in the English composition classes into which they are placed?

To be included in the analyses investigating the relationship of international students' scores on the individual sections of the iBT and grades, records needed to contain both a grade in an English composition class taken during the semester of admission and a breakdown of iBT scores in the different modalities. Of the obtained records, 200 contained the necessary information for these variables.

Descriptive Statistics

The statistics for the distributions of scores for the four different sections of the iBT (reading, listening, writing, speaking) for the 200 records qualifying for inclusion in this part of the study are listed in Table 17.

Table 17. Descriptive statistics for iBT section scores

iBT Section	Class Placement	N	Mean	Median	Standard Deviation	Kurtosis	Skewness
Reading	150	30	22.83	23	4.34	-0.45	-0.39
Reading	101B	29	22.52	24	6.10	-0.63	-0.58
Reading	101C	97	20.68	20	4.66	0.58	-0.41
Reading	101D	44	24.07	25	3.57	-0.68	-0.56
Reading	Undergraduate	156	21.44	21	4.97	0.01	-0.38
Reading	**All Classes**	**200**	**22.02**	**22**	**4.81**	**0.09**	**-0.51**
Listening	150	30	21.83	21	4.33	-0.96	0.26
Listening	101B	29	20.76	22	5.40	-0.39	-0.36
Listening	101C	97	18.60	19	4.76	0.31	-0.13
Listening	101D	44	22.52	22.5	3.20	0.49	-0.29
Listening	Undergraduate	156	19.62	20	4.97	-0.03	-0.12
Listening	**All Classes**	**200**	**20.26**	**21**	**4.78**	**0.08**	**-0.30**
Writing	150	30	22.87	22	3.60	0.06	-0.36
Writing	101B	29	18.72	18	4.46	-0.96	0.08
Writing	101C	97	20.11	21	3.18	0.26	-0.08
Writing	101D	44	22.82	22	2.30	-0.51	0.57
Writing	Undergraduate	156	20.38	21	3.75	-0.16	-0.12
Writing	**All Classes**	**200**	**22.87**	**22**	**3.60**	**0.06**	**-0.36**
Speaking	150	30	20.23	20	3.23	-0.11	0.11
Speaking	101B	29	17.93	18	3.90	-0.22	0.15
Speaking	101C	97	18.59	19	2.72	0.40	-0.05
Speaking	101D	44	18.84	18	2.76	0.03	0.32
Speaking	Undergraduate	156	18.78	19	3.14	0.25	0.03
Speaking	**All Classes**	**200**	**18.80**	**19**	**3.05**	**0.23**	**0.07**

It is also necessary to re-calculate distributions of grades in English composition classes for the sample of 200 individuals used in the correlations of iBT section scores, as they represent a new sample taken from the 503 records used in the correlations of composite iBT scores

and the descriptive statistics may change as a result. These statistics are displayed in Table 18.

Table 18. Descriptive statistics for distributions of grades used in correlations with iBT section scores

English Class	N	Mean (GPA)	Median	Standard Deviation	Kurtosis	Skewness
150	30	2.99	3.00	0.81	-0.55	-0.40
101B	29	3.05	3.33	1.14	1.97	-1.57
101C	97	3.01	3.33	0.99	1.08	-1.26
101D	44	3.88	4.00	0.31	7.35	-2.79
Undergraduate	156	3.01	3.33	0.98	1.17	-1.24
All Classes	200	3.20	3.67	0.95	1.88	-1.47

The most noteworthy different in these figures (beyond the smaller counts) is that kurtosis and skewness are much lower for the distribution of grades for this subset of the same English 101D group used in the composite score correlations. The median grade of 4.00 and GPA of 3.88 nevertheless still indicate the vast majority of participants in English 101D got an A, making for a poor range of grades to correlate with scores.

Correlations: Reading and Listening

The correlations involving the reading and listening sections of the iBT, which test knowledge and comprehension of English rather than production of language (and therefore do not represent a radical departure from the format of the PBT) will be examined first.

Statistics for correlations of iBT reading and listening scores and grades in English composition classes are displayed aggregately and by term in Table 19.

Table 19. Spearman correlations of iBT reading and listening scores with grades in English composition classes

iBT Section	English Class	N	r_s	Overlap	*p*
Reading	150	30	0.098	0.96%	0.54
Reading	101B	29	-0.554	30.69%	< 0.01
Reading	101C	97	-0.034	0.12%	0.89
Reading	101D	44	-0.054	0.29%	0.11
Reading	Undergraduate	156	-0.130	1.69%	0.15
Reading	All Classes	200	0.004	0.00%	0.66
Listening	150	30	0.369	13.62%	0.04
Listening	101B	29	-0.311	9.67%	0.14
Listening	101C	97	-0.115	1.32%	0.33
Listening	101D	44	-0.262	6.86%	0.49
Listening	Undergraduate	156	-0.100	1.00%	0.29
Listening	All Classes	200	0.042	0.18%	0.34

The most obvious pattern that is evident while examining Table 20 is that most of these relationships are negative, which is contrary to what one would expect. The relationship between English 150 grades and listening is the one noteworthy exception to this trend, while the coefficients for English 150 and reading as well as combined coefficients for both reading and listening are positive but only marginally so. The tendency towards negative coefficients reflecting inverse relationships is otherwise puzzling.

Since most of these figures are also not statistically significant, these mostly weak negative relationships are more likely evidence that reading and listening skills (at least as measured by these sections of the iBT) do not play a significant role in determining students' performance in these classes. The particularly glaring -0.554 coefficient for iBT reading and grades in English 101B represents a strong enough inverse relationship to be statistically significant. It is still unlikely that this finding reflects a situation where mastery of reading skills was actually detrimental to academic performance. Since the iBT was already taken before these participants were placed in classes, no hypothesis can be made that students were so busy studying for the iBT that they were not attending classes or otherwise not focused on their coursework (which could easily have been the case if these participants had completed these classes before taking the iBT). One possible explanation is that the

relationship is highly variable and that a different sample may have exhibited a much different, possibly even positive relationship, in light of the fluctuation between positive and negative relationships from one term to the next in several cases when composite iBT scores were correlated with grades.

Correlations: Writing and Speaking

The strength of the relationships between grades in classes and sections of the iBT that assess the production of language and integrative use of skills will now be examined. Statistics for correlations of iBT writing and speaking scores and grades in English composition classes are displayed aggregately and by term in Table 20.

Table 20. Spearman correlations of iBT writing and speaking scores and grades in English composition classes

iBT Section	English Class	N	r_s	Overlap	*p*
Writing	150	30	0.322	10.37%	0.07
Writing	101B	29	0.440	19.36%	0.02
Writing	101C	97	0.035	0.12%	0.56
Writing	101D	44	0.210	4.41%	< 0.01
Writing	Undergraduate	156	0.137	1.88%	0.05
Writing	All Classes	200	0.257	6.60%	< 0.01
Speaking	150	30	0.278	7.73%	0.11
Speaking	101B	29	0.624	38.94%	< 0.01
Speaking	101C	97	0.112	1.25%	0.19
Speaking	101D	44	-0.065	0.42%	0.10
Speaking	Undergraduate	156	0.228	5.20%	< 0.01
Speaking	All Classes	200	0.174	3.03%	< 0.01

The most noteworthy finding while investigating the strength of the relationships between iBT section scores and performance in English classes is that there is a consistently positive relationship between grades and production iBT scores, while mostly inverse relationships that were noted when the sections testing only comprehension and knowledge were correlated with the same grades. Reading correlates slightly more positively with grades

than speaking for participants who took English 101D, but in this both coefficients indicate very close to zero to begin with. For all other English composition classes (and when English classes are grouped), the correlations with grades for both speaking and writing scores were more positive than the correlation of grades of scores on both listening and reading.

Many of these observed positive relationships are nevertheless weak and represent a less than 10% overlap in abilities that determine the two variables. Interestingly, the strongest positive relationships for writing and speaking are observed of the same English 101B group for which the most strongly negative relationships for reading and listening were noted. The 0.624 coefficient for the correlation of iBT speaking and grades in English 101B, representing a 38.94% overlap between the two variables, is the strongest statistically significant relationship noted in this study. This is counter-intuitive considering that English 101B, a class focused primarily on grammar and structural issues in writing, is theoretically the least speaking-intensive of all the classes involved in the study. The coefficient of 0.440 for the correlation of iBT writing and grades in English 101B also stands out among these results. However, if the (relatively) strongly negative relationships noted for reading and listening when correlated with grades in English 101B owe to the unpredictable nature of the relationship between test scores in grades in these classes, then it is equally possible that a different sample may produce much different results in the coefficients for writing and speaking as well.

Summary

In correlations of scores on individual sections of the iBT with grades in English composition classes, the relationships once again vary considerably in their strength from one group to the next. Weak positive relationships are noted whenever scores representing each of the four skills are correlated with grades in all four levels of classes combined. What is particularly noteworthy is that in the case of the two iBT sections that evaluate the production of language, writing and speaking, the relationships with grades tend to be positive, while usually negative relationships of varying strength are noted when iBT reading and listening scores are correlated with the same grades. Meanwhile, positive relationships of varying strength are noted when the iBT writing and speaking sections are

correlated with grades in the same classes. Overall, a pattern is evident in which the writing and speaking sections of the iBT, the sections involving the production of language, are better indicators of academic performance than are the reading and listening sections.

CHAPTER 5
CONCLUSION

Because the new internet-based version of the TOEFL tests examinees' ability to produce spoken and written English and features tasks requiring the integrative use of multiple skills, it is theoretically an improved test of an international applicant's mastery of the most important skills for academic success. iBT scores now represent more than the comprehension and knowledge of structure that are assessed by older tests such as the PBT. However, in light of this study's findings, the extent to which the mastery of skills as measured by iBT scores can be considered an indicator of future success in coursework remains unclear. Depending on how participants in this study are grouped, relationships between TOEFL scores and academic success in the form of grades in English classes range from moderately strong to negligible to surprisingly negative. Meanwhile, Nevertheless, two key observations have been made that could support a hypothesis that the sections of the iBT concerning the production of language have positively contributed to the usefulness of the iBT as an indicator of actual proficiency as well as knowledge:

1. Stronger relationships (albeit not always tellingly strong in their own right) were noted in nearly every case in which the relationship between grades and sections of the iBT involving the production of language and integrative use of skills (writing and speaking), as compared to those more focused on comprehension alone (reading and listening).
2. At 0.324, the aggregate strength of the relationship between iBT composite scores and grades in English 150 is at least significantly higher than the relationships in earlier studies cited by Simner (1998) and Graham (1987), where coefficients less than 0.200 were usually noted. This is noteworthy because English 150, a non-ESL class, may better reflect the environment and expectations of the other non-ESL classes that international students must successfully complete in their programs of study. Furthermore, tasks in English 150 (group activities requiring cooperation with native-speakers and presentations as well as writing essays) require communicative competence in a wide variety of capacities, which the iBT is designed to better assess

than purely selected-response tests did. An argument could be made that iBT scores, though not dependable for predicting success in all situations, are more effective as indicators of potential when the production of language and the integrative use of all four skills (reading, listening, writing, and speaking) is particularly important.

This study therefore offers some evidence that the assessment of the production of language does indeed produce a stronger relationship between scores and success in practical use of the language than when knowledge and comprehension alone are tested. However, due to the overall inconsistency with which iBT scores forecasted performance in English classes taken during the semester of admission, this study does not lend support to the idea that TOEFL scores should be used as a basis for deciding whether international applicants possess the skills necessary to succeed. Should increasing numbers of international applicants prompt Iowa State University to become more selective, increasing the minimum required TOEFL score would not necessarily narrow the pool of qualified applicants down to those who are best prepared to succeed. Though TOEFL scores do provide institutions with useful information about international applicants' abilities, performance in previous coursework, especially if that coursework has been done in English, is likely to be a far stronger indicator of which applicants are best prepared (Graham, 1987; Simner, 1998). ETS' own warnings against use of the TOEFL as a sole criterion for admission decisions should be heeded in any case. It may still behoove admissions offices in universities throughout North America, as Simner (1998) suggested, to review policies where applicants are required to attain a minimum score on a proficiency test in order to first be considered for admission before other criteria are taken into consideration. When otherwise well-qualified international applicants nevertheless fall short of the minimum required score on a proficiency test and are denied admission on account of it, that in effect turns the proficiency test score into the sole criterion that ETS has warned against.

Meanwhile, the positive but only modestly strong relationships that have been noted between iBT scores and performance on ISU's English Placement Test. While these relationships are evidence that both tests are testing the same skills to a degree, the expansion of the role of the iBT for class placement purposes does not seem to be warranted based on

the findings of this study. Provided that the EPT is well-calibrated to the Department of English's needs for placement purposes, this test should continue to serve as the primary basis for placement decisions. The current policy of exempting incoming international students from the EPT if they have very high iBT scores may also be questionable in light of the frequency in this study with which examinees with high scores on the iBT nevertheless scored low on the EPT.

Limitations Acknowledged

Assuming that the test scores and grades acquired for this study were accurately matched the correct individuals in all cases before they were released for use in this study, the amount of data that were available and the nature of the correlations of scores with grades in this study still necessitate some caution in their interpretation. Although international enrollment at ISU has been strong over the past years, and a large pool of international student records was available, the specific criteria that had to be met for inclusion in each correlation always meant that most of the available records had been ruled out when statistics were being calculated. Sample sizes were often large enough that findings in this study are statistically significant, but it is still risky to make generalizations about issues affecting the whole university or especially proficiency testing all over the whole world based only on these samples. Furthermore, taking correlation coefficients at face value whenever scores on proficiency tests are correlated with grades in different classes (or even different sections of the same class when there are different instructors for each section) assumes that proficiency will play the same role in determining a student's amount of success in each class. Although all of the classes involved in this study are English composition classes, each class is intended for a different proficiency level, and evaluation criteria is not likely to be the exactly the same even between two different sections of the same English class. The mixing of groups that was necessary to calculate correlations with larger sample sizes created situations where "apples to oranges" comparisons were arguably being made, and the resulting coefficients cannot safely be considered absolutely indicative of the real relationship as a

result. When groups were not mixed, the smaller sample sizes facilitated fluctuating coefficients from one group to the next that make it difficult to draw any conclusions about the overall nature of the relationship between variables. Most of all, even if the iBT could be proven beyond any doubt to be an excellent indicator of English proficiency, the role English proficiency in determining an international student's success or failure and the role proficiency test scores should therefore play in admissions decisions would nevertheless remain very much in question.

Suggestions for Further Research

The trend that emerged in this study that the speaking and writing sections of the TOEFL iBT correlate more strongly with grades in classes merits further investigation. If other studies discovered similar trends, this could contribute to discussion of the importance of production of language in language testing. Studies involving larger samples than were available for correlations involving iBT section scores in this study would serve as a better basis for generalizing about the relationship between mastery of language production and academic success. If ISU continues to keep records of international students' breakdown of iBT scores into the four different modalities, in a few years it will be possible for studies similar to this one to be carried out in which a far greater number of records is included in the correlations.

In light of the greater consistency with which positive relationships between iBT scores and grades in English 150 were noted, further investigation of how iBT scores correlate with success in other non-ESL classes may be productive as well. In addition to further studies of involving non-ESL English classes with larger sample sizes, studies could examine the relationship between iBT scores and grades in non-ESL classes in other fields such as other social sciences, physical and life sciences, and business. Though English classes were used in this study as examples of English language-intensive academic environments where the English proficiency would make the greatest difference, the ultimate

degree of success of most international students at Iowa State University will be determined by their performance in non-ESL classes in other fields.

REFERENCES

Alderson, J., Krahnke, K., & Stansfield, C. (Eds.). (1987). *Reviews of English Language Proficiency Tests.* Teachers of English to Speakers of Other Languages: Washington, DC, USA.

Bachman, L., & Palmer, R. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Bachman, L. (2004). Statistical Analyses for Language Assessment. Cambridge, UK: Cambridge University Press.

Bailey, K. (1999). Washback in Language Testing. *TOEFL Monograph Series*, *15* (June 1999)

Bannerjee, J. (2003). The TOEFL CBT (Computer-based test). *Language Testing, 20*, 111-123.

Brown, J. (2003). Standardized Tests, Classroom Tests, and the Case for On-Campus ESL Testing. In Douglas, D. (ed.). *English language testing in U.S colleges and universities,* (2nd ed.). Washington, D.C: Association of International Educators.

Chalhouh-Deville, M. (2003). Fundamentals of Admissions Tests: MELAB, IELTS, and TOEFL. In Douglas, D. (Ed.). *English language testing in U.S colleges and universities,* (2nd ed.). Washington, D.C: Association of International Educators.

Chen, P. & Popovich, P. (2002). *Correlation: Parametric and Nonparametric Measures*. Thousand Oaks, CA: Sage Publications.

Douglas, D. (Ed.). (2003). *English language testing in U.S colleges and universities,* (2nd ed.). Washington, D.C: Association of International Educators.

Douglas, D. (2010). *Understanding Language Assessment.* London, UK: Hodder.

Educational Testing Service. (2011a). TOEFL Home. Retrieved on July 17th, 2011 from http://www.ets.org/toefl/

Educational Testing Service. (2011b). Internet-based Test: Test Content. Retrieved on June 15th, 2011 from http://www.ets.org/toefl/ibt/about/content

Educational Testing Service. (2011c). Paper-based Test: Test Content and Structure. Retrieved on June 15th, 2011 from http://www.ets.org/toefl/pbt/about/content

Enright, M., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., et al. (2008). Prototyping New Assessment Tasks. In Chapelle, C., Enright, M., & Jamison, J. (Eds.). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY/London, UK: Routledge.

Graham, J.. (1987). English Language Proficiency and the Prediction of Academic Success. *TESOL Quarterly, 21(3),* 505-521.

Hamp-Lyons, L. (1997.) Washback, impact and validity: ethical concerns. *Language Testing, 14,* 295-303.

Hamp-Lyons, L. (1998.) Ethical Test Preparation Practice: The Case of the TOEFL. *TESOL Quarterly, 32(2),* 327-339.

Hamp-Lyons, L., & Shohamy, E. (2003.) *The Effect of the Changes in the New TOEFL Format on the Teaching and Learning of EFL/ESL: Stage 1 (2001-2003): Instrument Development and Validation.* University of Melbourne.

Inside Higher Ed. (2008). New Challenge to the TOEFL. Retrieved on July 16th, 2011 from http://www.insidehighered.com/news/2008/10/02/english

Iowa State University. (2011a). ESL Courses. Retrieved on July 16th, 2011 from http://www.public.iastate.edu/~apling/engl101.html

Iowa State University. (2011b). English Placement Test - FAQ. Retrieved on June 15th, 2011 from http://www.public.iastate.edu/~apling/ept.html

Livingston, S. (2009). Constructed Response Test Questions: How We Use Them; How We Score Them. *ETS R&D Connections*, *11* (November 2009)

Karabulut, A. (2007). *Micro level impacts of foreign language test (university entrance examination) in Turkey: a washback study*. (Master's thesis) Iowa State University, Ames, Iowa.

Simner, M. (1998.) Use of the TOEFL as a Standard for University Admission. *European Journal of Psychological Assessment*, 14(3), 261.

Stevenson, D. (1987). Test of English as a Foreign Language. In Alderson, J., Krahnke, K., & Stansfield, C. (Eds.). *Reviews of English Language Proficiency Tests.* Teachers of English to Speakers of Other Languages: Washington, DC, USA.

Taylor, C. & Angelis, P. (2008). The Evolution of the TOEFL. In Chapelle, C., Enright, M., & Jamison, J. (Eds.). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY/London, UK: Routledge.

Wall, D. and Horák, T. (2006). The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 1, The Baseline Study. ETS, *TOEFL Monograph Series, 34* (June 2006)

Wang, L., Eignor, D., & Enright, M. (2008). A Final Analysis. In Chapelle, C., Enright, M., & Jamison, J. (Eds.). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY/London, UK: Routledge.

Wessa, Patrick. (2011), Free Statistics Software, Office for Research Development and Education. Retrieved May/June 2011 from http://www.wessa.net/rankcorr.wasp

ACKNOWLEDGEMENTS

I would like to personally thank Volker Hegelheimer and Yoo Ree Chung of the Department of English as well as Judy Minnick and Jonathan Compton of the Office of the Registrar for their willingness to help me obtain the data necessary to carry out this study.