The CATESOL Journal

DANIEL MOGLEN
University of California, Davis

The Re-Placement Test: Using TOEFL for Purposes of Placement

This article will consider using TOEFL scores for purposes of placement and advising for international graduate students at a northern California research university. As the number of international students is on the rise and the funds for the graduate ESL program are diminishing, the way in which the university is handling the influx of international students is undergoing substantial changes. One aspect of the system that is gaining attention is the graduate-level ESL placement exam. To find out if using TOEFL scores for placement is a viable option, I have looked at the Pearson *r* coefficient for TOEFL scores and university placement exam scores from years 2007-2011. Results from this study show a moderate correlation between the TOEFL and placement exam and suggest that students at this university with TOEFL scores 110 and above should be exempt from any ESL requirement while students with TOEFL scores below 90 need to take ESL courses.

Introduction

The process of language testing, assessment, and placement has been closely examined and widely researched (Brown, 1989; Kelleher, 2008; Raimes, 1990). As testing and assessment are grounded upon the idea of validity, fairness, and reliability, I will take these factors into account in examining the TOEFL iBT (Test of English as a Foreign Language Internet Based Test) and the graduate-level ESL placement exam at a northern California university. The former is a standardized test of monolithic proportion that determines the academic fate of innumerable international students who are pursuing an education in America, at both the undergraduate and graduate levels. The graduate-level ESL placement exam (known as "placement exam" from this point) is created at the university department level and serves as a filter to determine which newly admitted international students are required to take an ESL course, and if so, which one.

The number of international graduate students attending universities in the US has been steadily increasing over the last decade, and in 2014 the number was nearly 330,000 (Institute of International Education, 2015). This northern California university is no exception, and the increasing number of international students, coupled with the decreasing amount of available funds, is resulting in major changes in the graduate-level ESL program. As a result, the placement exam and the placement process are being reevaluated. According to the university's catalog, students who are flagged as needing to take the placement exam are those for whom English is not their native language nor their language of previous instruction. As it stands now, these students must take the placement exam in the weeks before the beginning of the fall quarter. In this time-sensitive period of shuffling around schedules and getting acquainted with a new lifestyle, each student must be prepared for potentially not passing the placement exam and having to sign up for an extra class. Some students strategically sign up for the graduate-level ESL class preemptively, which results in classes that are overenrolled and with students placed on the wait list. This all makes for a hectic, often confusing beginning to the academic year. In this article, the literature review will address the interpretation of the TOEFL scores, its history, format, and characteristics. I will then address the following research questions:

1. How well do TOEFL scores correlate with the ESL placement exam scores?
2. How well do the TOEFL subset scores correlate with the placement exam?
3. How well do the placement exam subset scores correlate with the total ESL placement exam?
4. How well do the ESL placement subset scores correlate with each other?
5. How well do the subset scores of the TOEFL correlate with the subset scores of the placement exam?

I hope that answers to these questions will aid in making recommendations about the future structure of the graduate-level ESL placement exam, as well as influence the way that the university handles the placement process.

How to Interpret TOEFL Scores

Scores are the end result for both the test taker and the institution, or "score consumer" (Fulcher & Davidson, 2007), and these numbers are the primary evidence of the test taker's language proficiency. The

institution does not see any qualitative performance, but rather must render a decision based on the quantitative output. An aggregate score of 80, which is the minimum accepted TOEFL score at the northern California university, could represent varied levels in each skill set. While one student may be proficient in reading and listening yet poor in speaking and writing, another may be proficient in speaking and writing yet poor in reading and listening. This is also the case regarding oral speaking proficiency, for which Iwashita, Brown, McNamara, and O'Hagan (2008) note that "speakers may produce qualitatively quite different performances and yet receive similar ratings" (p. 27). In this way, the quantitative score, with no qualitative input, masks how the test taker achieved the score. Consequently, students who share the same total score could potentially have vastly different linguistic skill sets.

Notably, more than 8,500 colleges and universities use TOEFL scores for admissions (Educational Testing Service/ETS, 2015), and each institution is charged with making sense of how to use them. ETS states on its website, www.ets.org, that admission decisions should not be based solely on the TOEFL score, but rather it should be used in conjunction with other criteria. Research in the field generally corroborates this stance. For example, Wongtrirat (2010) conducted a meta-analysis of the predictive qualities of TOEFL, in particular its ability to predict academic achievement, defined as "GPA, numbers of courses completed, or both" (p. 14). The results of the analysis showed that TOEFL scores had a low positive correlation with academic achievement, including findings from Ng (2007), who looked at 433 international students, as well as from Krausz, Schiff, Schiff, and Van Hise (2005) and Zhang (1996). Rarely was a high correlation reported, and when it was, as in the study by Burgess and Greis (1970), the results were not generalizable because of the small sample size. And, it is unclear how the results of their study would translate today given the evolution of the TOEFL throughout the years since the 1970s. Interestingly, the minimum accepted TOEFL score for a particular institution can affect the predictability of the TOEFL score. An institution that requires a high TOEFL score will have a smaller range of scores, while a lower minimum score will have a larger range. Because in both cases there is a wide range of academic achievement, a wider range of TOEFL scores will yield a higher correlation with academic achievement (Wongtrirat, 2010).

A History and Overview of the TOEFL

Since the TOEFL's inception in the early 1960s, the test has had more than 25 million test takers worldwide, and it has test centers

in 165 countries (www.ets.org). The administering organization, ETS (Educational Testing Services), describes the test:

> The TOEFL test measures a student's ability to use and understand English at the university level and evaluates how a student combines reading, speaking, listening, and writing skills needed to succeed in an academic setting. (www.ets.org)

The themes in the literature revolving around TOEFL take into account history (Spolsky, 2007), fairness (Kunnan, 2010; McNamara & Roever, 2006), and validity (Xi, 2010). The history of the TOEFL provides the context and background, while fairness and validity are two contested issues that represent some of the ongoing debates in the literature of language assessment. Perspectives vary about the TOEFL; however, there is no dispute about the immense influence that this test wields in the admissions process for countless international students.

The test has undergone a series of developments throughout its history (see Table 1), evolving from a paper-based test to a computer-based test, and finally, in 2005, the iBT, or Internet-based test. The challenges of the test's evolution are significant, namely maintaining a high level of standardization and thus reliability. In each stage, the format and content must translate into an accurate scoring system, which Spolsky (2007) likens to navigating a "modern supertanker" (p. 14). Specifically, each test must be calibrated to previous tests, dating all the way back to the first one.

Not only has the format changed, but the constructs and the content have also evolved. At the inception of the TOEFL, the constructs focused more on "discrete components" (Table 1), achieving this through multiple-choice questions. In the second phase, in addition to the original constructs, writing and speaking skills were added, and the modality changed from a paper- to a computer-based test. The current format of the TOEFL operationalizes communicative competence (www.ets.org), which includes both grammatical and pragmatic competence (Savignon, 1991), through the integration of multiple skill sets in its four sections, speaking, reading, listening, and writing, which are described below. Each section is scored out of 30 points, for a maximum score of 120.

The fact that the TOEFL has undergone a change in modality, from paper to computer, deserves attention. In its newest state the TOEFL is a computer-adaptive test (CAT), meaning that the difficulty of the questions adjusts to the performance of the test taker. In other words, the more questions that the test taker answers correctly, the harder the questions become, or vice versa.

Table 1
History of TOEFL

Stages	*Construct*	*Content*
1. The first TOEFL Test 1964–1979	Discrete components of language skills and knowledge	Multiple-choice items assessing vocabulary, reading comprehension, listening comprehension, knowledge of correct English structure and grammar
2. A suite of TOEFL Tests 1979–2005	Original constructs (listening, reading, structure, and grammar) retained with two added—writing ability and speaking ability	In addition to multiple-choice items assessing the original constructs, separate constructed-response tests of writing, the TWE test and speaking, the TSE test, w,ere developed
3. The TOEFL iBT Test 2005–present	Communicative competence—the ability to put language knowledge to use in relevant contexts	Academic tasks were developed that require the integration of receptive and productive skills such as listening, reading, and writing or speaking, as well as multiple-choice items for listening and reading

Note. *Adapted from* TOEFL Program History, *p. 4. TOEFL iBT Research Insight, Series 1, Vol. 6. Educational Testing Service. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf*

Because the test is administered around the world, there is undoubtedly a discrepancy among test takers in terms of computer familiarity and experience. Dooey (2008) studied test anxiety and found through questionnaires that in comparison to previous versions of the TOEFL, "examinees appeared to become more comfortable with using the computerized TOEFL test" (p. 28). Also, English ability and experience with computers did not have an impact on test anxiety level (Dooey, 2008). This is good news, and as computers continue to grow in their prevalence as an educational tool, test takers' anxiety level due to the computer modality will surely decrease.

Central to a test's characteristics, especially one of this magnitude, must be fairness. Fairness is defined as an "absence of bias, equitable treatment of all test takers in the testing process, and equity in opportunity to learn the material in an achievement test" (Standards

for Educational and Psychological Testing, as described in Xi, 2010, p. 147). In this way, fairness intersects validity, access, and justice. Xi (2010) posits a fairness argument based within a validity argument, saying that "anything that weakens fairness compromises the validity of test score interpretation and use" (p. 154).

The test takers play a significant role as well, so it would be prudent to look at the test from their perspective. The literature acknowledges that the TOEFL provides the same test for everyone, in spite of the fact that the test takers make up an incredibly diverse group of people, both culturally and linguistically (McNamara & Roever, 2006). Therefore, special attention needs to be paid to ensure that certain cultural/linguistic groups do not have an unfair advantage, and that bias is minimized. The reality is, however, that the institutions hold a great amount of power, and in the words of McNamara and Roever (2006), "as long as the score users are satisfied with the test's fairness, the instrument's acceptability does not suffer and test makers can expect few repercussions" (p. 136).

Bias is defined by McNamara and Roever (2006) as "a skewed and unfair inclination toward one side (group, population) to the detriment of another" (p. 82). In assessment, however, the more neutral term "differential item functioning" (DIF) is used in analysis to avoid the connotations of "bias." For example, a test would not be considered to have bias if there were some vocabulary items that had cognates in another language because that may reflect real-world usage of the language (McNamara & Roever, 2006).

TOEFL Subsections

Reading Section. The first section that the test taker encounters is the reading section. The section has a time limit of 60-80 minutes, has 36-56 questions, and requires the test taker to respond to questions after reading academic passages (www.ets.org). The reading section contains 3-4 excerpts that are about 700 words long, which is an increase from previous versions of the test, which used excerpts of about 350 words. This increase was based on the "rationale that longer passages can better approximate the academic reading load at North American universities" (Liu, 2011, p. 236).

Liu (2011) used DIF to analyze the effect, if any, of academic major and cultural familiarity on the reading-section performance of the TOEFL iBT. Liu (2011) gathered data from 8,692 participants in the form of a questionnaire to identify their major and cultural familiarity. Then they were divided into the focal group, participants with relevant expertise, and the reference group, which included everyone else. Although there were certain individual questions that yielded a

difference among the groups, the diversity of question types made any performance difference negligible when measuring performance on the passage level and across a variety of passages (Liu, 2011).

Listening Section. The listening section is 60-90 minutes and has 34-51 questions that require the test taker to respond to questions after listening to a conversation or lecture (www.ets.org). The listening passage appears before the questions, and so the test taker is required to try to remember everything. Although note taking is allowed, the test taker is unaware of the test questions beforehand. An example of a question from the listening section comes from the ETS website and is shown here.

1. You will hear:	
(man)	*Shall I lock up the computer lab now before I go home?*
(woman)	*Don't bother. I'm not leaving for a while, I can check it on my way out.*
(narrator)	*What will the woman probably do?*
You will read:	*A. Lock the computer lab later.*
	B. Leave with the man.
	C. Buy a new lock for the computer lab.
	D. Show the man where the lab is.

(www.ets.org)

This question requires the test taker to both understand and to make an inference based on the short dialogue. The content of the question can be highly culture specific, and this could lead to construct irrelevance in that the test taker must hold some cultural/pragmatic knowledge in order to correctly answer the question. The next part of the listening test requires the test taker to listen to longer dialogues/monologues and then answer multiple questions.

Writing Section. The time limit is 50 minutes for the writing section of the TOEFL, in which the test taker has to complete two tasks. The writing section tasks may include a response to a reading or a prompt or may require the test taker to express an opinion (www.ets.org). This is the final section of the TOEFL. Enright and Quilan (2010) state about the writing section of the TOEFL:

> Given the time constraints inherent in the testing situation, test takers do not have much time to revise and polish their writing. Nevertheless, these timed writing exercises are sufficient to pro-

> vide evidence of examinees' basic writing skills, from facility of language use to the organization and development of ideas. (p. 319)

This time constraint cannot be underestimated, as L2 writers commonly implement different strategies when writing and may need more time to complete a writing task in comparison to writers using their L1 (Raimes, 1990; Silva, 1993). Students who have honed their timed-writing skills have a great advantage in this test setting. As the TOEFL is concerned with testing for preparedness in the academic English setting, it is important to note that this section requires writing something equivalent to a first draft, and it does not allow enough time for ample editing and revision. The timed-writing skill remains useful in academia in terms of test taking.

The content of the writing prompts is another point of consideration. Given the diversity of the test takers, what kinds of prompts would be suitable? An example of a writing topic is:

> It has recently been announced that a new movie theater may be built in your neighborhood. Do you support or oppose this plan? Why? Use specific reasons and details to support your answer. (www.ets.org)

In the writing section, test takers are not offered options, and they must write on the topic that appears, regardless of their breadth of knowledge or amount of interest about the particular topic.

Speaking Section. The speaking section lasts 20 minutes and includes six tasks (www.ets.org). For example, the test taker may encounter:

> Talk about a pleasant and memorable event that happened while you were in school. Explain why this event brings back fond memories.
> **Preparation Time: 15 seconds Response Time: 45 seconds**
> (www.ets.org)

Other speaking tasks include listening to a conversation and summarizing it as well as reading a passage and talking about the main points. These exercises allow 20 seconds for preparation and 60 seconds for the response, and note taking is allowed.

Papajohn (2006) notes that the maximum speaking time on the speaking section is 5.5 minutes. This can be seen only as a sample of the test taker's ability and not necessarily representative of his or her communicative competence. This is especially relevant for the

graduate students who intend to become teaching assistants, or TAs. Because one of the primary duties for TAs is to lead a weekly discussion section, one must be equipped with enough linguistic skills and resources to present academic material, field questions from students, and manage classroom behaviors. The current policy requires that an international graduate student must receive a score of 23 or higher to be considered for a TA position. The policy also states that although there are some exceptions and waivers granted, the absolute minimum is a TOEFL speaking-section score of 20. Additionally, this requirement may be met with an IELTS speaking score of 7 or a SPEAK test score of 45 (ITA policy).

An Overview of the ESL Placement Exam

The graduate-level ESL placement exam at the northern California university was developed during the course of two decades. Earlier versions of the exam included a reading and grammar section, but those were discarded without any significant loss to the test's utility. In its current state, there are three sections. The first part is the cloze test, in which a short article, or excerpt of an article, is displayed with 20 words missing. The other two parts require a written response: one, in the form of a response to a prompt, and the second, in the form of a summary in response to a live lecture. The highest score is 100, and a score of 70 or above represents a passing score, meaning that no ESL course is required. A score of 60-69 means that the student should take one ESL course, and 59 or lower suggests that the student should take two ESL courses. In 2011, a new policy was instituted that exempted students with a TOEFL score of 110 or higher from taking the placement exam and from any ESL requirement.

Cloze Test

For the cloze test, there are 20 blanks with each blank replacing a word from an article excerpt. Each blank is worth one point for a total of 20 points. An example of the cloze test is the following:

> Many animals in captivity have produced drawings spontaneously, and many more have been _____1_____ to draw or paint by trainers or other humans. For instance, Siri, an Indian elephant, was often seen _____2_____ scratches on the floor of her enclosure with a pebble. When Siri's trainer provided her with drawing _____3_____, she responded _____4_____ producing dozens of "pictures." Two artists, who commented on the drawings _____5_____ knowing who had produced them, admired their "flair and decisiveness and originality."

Some acceptable answers for (1) include "taught" and "trained" while "teach" or "teached" were marked as incorrect. For (2), the graders accepted "making" or "drawing" but not "doing." See Table 2 for correct and incorrect responses.

Table 2
Correct and Incorrect Responses

Example	*Correct responses*	*Incorrect responses*	*Form tested*
1	taught, trained	teach, teached	Present perfect
2	making, drawing	doing	Gerund (-ing)
3	pencils, colors, tools	tool	Plural –s
4	By	through, that, immediately	Preposition
5	Without	pictures	Preposition

The grading process in 2011 (previous years had similar grading procedures) consisted of a group of 10 people, seven graduate students and three faculty members, who individually completed the test filling in all of the words that they thought were acceptable. The group then went over each answer and decided what would be acceptable or not. Once everyone had agreed, the graders marked the tests, and questionable answers were discussed throughout the process. Attention was not given to spelling, as long as the word was recognizable.

Essay Response to a Prompt

The next part of the test is the essay, which was also scored out of 20 points. In this section, the test taker had to choose one of three essay topics and write on that topic. A score of 14-20 was considered passing, with the delineation of 14-16 as adequate and 17-20 as strong. Papers that scored 13 and under were considered not passing, with 9-13 scoring in the weak range, and 0-8 fell in the extremely weak range (see Table 3).

Table 3
Essay Scoring Categories

Essay scoring categories	*Score*
Strong	17-20
Adequate	14-16
Weak	9-13
Extremely weak	0-8

The directions for the essay section were as follows:

> Directions: Write a response to ONE of the following topics. Your writing will be scored for organization, development of ideas, coherence, and language control. Whichever question you choose, make sure you support your response with specific details and reasoning.

An example prompt is:

> The pace of life and how people view time and punctuality can vary from culture to culture. Compare the pace of life and people's attention to time in your culture of origin and American culture in order to determine to what extent they are similar or different. Then comment on which pace of life you, personally, are most comfortable with and why.

The prompts tended to elicit culturally themed or opinion responses, in which there was no particular need to use specialized vocabulary. This had the effect of leveling the playing field, as the students who are taking the test come from a wide range of academic disciplines. Also, the fact that there are three prompts to choose from gives the student the chance to pick the topic that he or she feels most comfortable or confident with. This may have a countereffect if the student has trouble deciding or is left with the feeling of having chosen the wrong topic, especially under time pressure.

Summary Response to a Lecture

The final part of the test consists of a live lecture and a corresponding handout. The test takers need to take notes during the lecture and then write a summary of the lecture. This is supposed to mimic a skill that is necessary to survive in graduate school, namely, listening to a lecture and synthesizing the material. The lecture section of the placement exam focused on topics that would be of interest for international graduate students, but also special attention was placed to ensure that the topics would be obscure enough that the test takers would likely not have much prior knowledge on the topic. Some examples include a lecture on the "Mozart effect" or Lake Tahoe water quality.

The summary is given the most weight and is scored out of 30. Scores of 21 and over represent passing papers, with 21-25 signaling a clearly adequate paper, and 26-30 a strong paper. Papers that score 20 or less are considered not passing, with 16-20 less than adequate, 11-15 weak, and 0-10 extremely weak. See Table 4 for a display of the scores and scoring categories.

Table 4
Summary Scoring Categories

Summary scoring categories	*Score*
Strong	26-30
Clearly adequate	21-25
Less than adequate	16-20
Weak	11-15
Extremely weak	0-10

The instructions were as follows:

> On the paper provided, take notes while you listen, then write, a summary. You have 40 minutes after the lecture to complete your summary.

While the essay was more open ended, the summary was intended to elicit a more rigid response. Instead of expressing an opinion, or sharing an experience, the test taker had to summarize and synthesize the lecture. Papers that lacked key information from the lecture were marked down accordingly.

Grading Process for the Essay and Summary

A norming process for both the essay and summary sections ensured reader reliability. This process started with two faculty members' reading through the exams to find essays/summaries that exemplified each scoring category. They also packaged three "scrambled sets," one essay from each scoring category in random order. All of the test graders met, and the faculty members led the discussion explaining the criteria for each scoring category. For example, the criteria for a quality essay included complexity of sentences, flow of ideas, and organization. Spelling and various local errors, however, were not considered as decisive in terms of affecting the overall score. Some essays had very few errors but still scored poorly because of overly simplistic vocabulary and sentence structure.

The graders then read an example from each category and discussed why it was scored in that way. After that, they were given the first scrambled set to score, and the tests were placed in the appropriate scoring category. A discussion followed each scrambled set to ensure that all of the graders were normed. Only after going through three scrambled sets were the graders ready to read the essays individually. Experienced graders checked scores, and papers with borderline scores were reread.

Methodology

Data Collection

TOEFL scores were obtained from the Office of Graduate Studies, as was other student information such as name, country of origin, degree, and degree objective. Because all three types of TOEFL tests (paper-based, computer-based, and Internet-based) were present in the data, the researcher used the conversion chart found on the official ETS website to synchronize the scores (http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Comparison_Tables.pdf). The TOEFL iBT test scores were left unchanged, while the computer-based and paper-based scores were converted.

Participants

In the years 2007-2011, 767 international graduate students took the ESL placement exam. This population represents 46 countries, with the top three countries being China, Korea, and Taiwan. The graduate programs varied even more widely than the countries, with 63 graduate programs reported. The most popular programs for these international graduate students include Electrical and Computer Engineering, Civil and Environmental Engineering, Computer Science, and Chemistry. TOEFL subset and total scores for the population are displayed in Table 5.

Table 5
TOEFL Subset and Total Scores

	Average	*Standard deviation*	*Range*
Listening	24.20	4.37	2-30
Reading	25.94	3.94	5-30
Speaking	20.42	3.29	12-30
Writing	23.57	3.49	12-30
Total	94.56	11.54	35-118

Notably, the average score between 2007 and 2011 was 94.56 with a standard deviation of 11.54. Because the minimum TOEFL score for admissions is set at 80, this means that students whose score was one standard deviation below the mean still met the minimum TOEFL score requirement. Students were the weakest on the speaking section, which interestingly is not taken into consideration for placement purposes based on the current state of the placement exam. That said,

the speaking score plays a role in TA hiring decisions. The reading and listening sections were negatively skewed (see the Appendix for histograms), showing a denser concentration of high scores. Perhaps this is a result of successful test preparation.

The averages for the placement exam (see Table 6) were comparatively lower. The average total score was 58.78, with a standard deviation of 12.96. That means that students who scored one standard deviation above the mean were on the threshold of passing. This could be a direct result of the test taker's inability to prepare for the placement exam because the test content is not disclosed.

Table 6
ESL Placement Exam Scores

	Average	*Standard deviation*	*Range*
Cloze test/20	12.01	3.56	2-20
Summary/30	17.10	4.53	0-29
Essay/20	12.17	2.71	5-21
Total/100	58.78	12.96	20-94

Table 7 shows grouped placement exam scores that correspond to placements. Just under one-quarter of the test takers from the years 2007-2011 passed the placement exam, while the other 73.7% of the population tested into an ESL course.

Table 7
Frequency and Percentage of Grouped Placement Exam Scores

	Frequency	*Percent*	*Cumulative percent*
No score reported	22	2.8	2.8
70 and above	185	23.5	26.3
60-69	205	26.0	52.3
59 and below	376	47.7	100.0
Total	788	100.0	

These tables are showing that an average-scoring student in the population scored well enough on the TOEFL for admission but did poorly enough to be required to take two ESL courses. So the question remains: How do we reconcile the reasonably high average of TOEFL scores, yet the less-than-adequate showing on the placement exam? Some implications of this will be discussed in the sections that follow.

Data Analysis

Figure 1 displays the population in groups based on their TOEFL scores and shows their corresponding mean placement-exam scores. This graph illustrates that the mean placement-exam score increases as the scoring category increases. Notably, the two scoring categories that average higher than a 70 on the placement exam are 110-114 and 115-120. This supports the current policy that exempts students who have TOEFL scores of 110 or above from taking the placement exam.

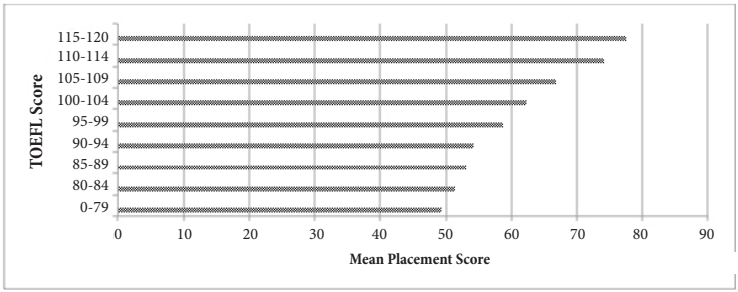


Figure 1. TOEFL score ranges and corresponding placement scores.

What this graph does not show, however, are the *N*'s for each category. These are displayed in Table 8. A mere eight students out of the population scored between 115 and 120, which represents a score in the 97th percentile out of all graduate students taking the TOEFL (www.ets.org). The scoring range with the highest *N* is 100-104, with 92.

Table 8
TOEFL Scoring Categories and Placement Exam Performance

TOEFL scores	*Mean placement score*	*Standard deviation*	N
0-79	49.33	11.02	45
80-84	51.39	10.37	56
85-89	53.10	11.67	60
90-94	54.21	9.69	82
95-99	58.69	11.58	75
100-104	62.39	10.96	92
105-109	66.75	10.72	60
110-114	74.14	12.51	36
115-120	77.50	11.14	8
Total	58.78	12.96	514

Let us revisit the original research questions, which are:

1. How well do TOEFL scores correlate with the ESL placement exam scores?
2. How well do the TOEFL subset scores correlate with the placement exam?
3. How well do the placement exam subset scores correlate with the total ESL placement-exam scores?
4. How well do the ESL placement subset scores correlate with each other?
5. How well do the subset scores of the TOEFL correlate with the subset scores of the placement exam?

In response to research question (1), the Pearson *(r)* correlation for TOEFL scores and placement exam scores is .52. This represents a moderate correlation, but because the total TOEFL score is a summation of the subset scores, it is necessary to address research question (2), which looks at the correlation for the TOEFL subset scores and the placement exam scores. The speaking section of the TOEFL represents the highest correlation with the placement exam (r=.56). The next highest is the writing section (r=.48), followed by the listening (r=.45), and finally reading (r=.29). It is an interesting finding that the speaking section is the most highly correlated part of the TOEFL with regard to the placement exam, especially because the placement exam does not measure speaking. One explanation could be that because the population's performance was the weakest in the speaking section, this section was the most reflective of language ability.

In response to research question (3), all of the placement subset scores were well correlated with the placement exam score. The cloze test and placement score correlation yielded r=.80, while the summary and placement score was r=.88, and the essay and placement was r=.80. Because all sections correlated with .80 or higher, this shows that there may be some redundancy on the placement exam.

The subset scores of the placement exam had a moderate correlation with each other, with the essay and summary sections ranking the highest (r=.61), followed by cloze test and essay (r=.53), and cloze test and summary (r=.50). The r value is the highest for the summary and essay parts of the placement exam; in other words, here lies the highest level of redundancy. The cloze test and summary are still moderately correlated, but they seem to be assessing different language skills. This finding, in conjunction with the results from research question (3), leads to the conclusion that the essay section is expendable without

much loss to the overall utility of the placement exam score. Further evidence for this is the high *r* value (.95) for the correlation between the cloze and summary subset scores and the total placement exam score.

For research question (5), the listening, reading, and writing sections of the TOEFL correlate the best with the cloze test of the placement (r=.46, r=.38, and r=.50 respectively), while the speaking section of the TOEFL correlates the best with the summary section of the placement exam (r=.48).

Limitations

One of the limitations of using TOEFL scores in this study was the fact that many students took the test multiple times, sometimes resulting in a wide array of scores. The researcher chose to use the highest score for the sake of consistency, but that shrouds the test taker's lower-scoring performances. One example, although it is an exception, is a dramatic case for the utility of the placement exam. This student took the TOEFL test five times, scoring as low as 47 and as high as 92. On the placement exam, however, this student scored a 27 (one of the lowest scores of the population). In other words, the TOEFL score was just under the population mean, while the placement-exam score was more than two standard deviations below the mean. Because one cannot prepare for the placement exam and must take it "blindly," perhaps it is more apt in showing the student's ability. Moreover, the university is merely given the TOEFL and subset scores without any qualitative evidence of linguistic performance. Meanwhile, we can look at the placement exam and see which errors were committed, how often, and so forth.

Another consideration is the group of students who scored relatively poorly on the TOEFL but performed well on the placement exam. One explanation for this is that these students might have taken an intensive English course in the time between the TOEFL and placement exam. In effect, the placement score is more reflective of their current language ability compared to their score on TOEFL, which is a test that they might have taken six months or a year before being admitted to university.

One further limitation was the fact that the database available for this study did not provide TOEFL scores for every student. As a result, statistical analyses including the correlations did not represent every incoming international graduate student, but rather those who took the placement exam and had reported a TOEFL score.

Discussion

Although there is a moderately positive correlation, there is no conclusive evidence that TOEFL scores can predict academic achievement or placement scores. This especially holds true for the middling scores, and less true for the extreme scores. As shown from the data, students who score 79 or lower on the TOEFL had an average score of 49.33 on the placement exam. Based on experience, this shows that further ESL instruction is needed for these students to be successful in their graduate programs at this university. Students with 110 or above TOEFL scores, on the other hand, averaged upward of 75 on the placement exam, and, with few exceptions, proved themselves to be ready for rigorous academic work.

A number of factors may account for this pattern, including the nature of the test, the institutionalization of the test, and the content of the test. With regard to the TOEFL, the test taker can take crash courses and has ample access to study guides and materials. In addition, it is perfectly acceptable to take the test multiple times, and the lower scores are disregarded in favor of the highest one. Furthermore, the TOEFL has established itself as a key, defining marker for admissions. Even though ETS warns that admission decisions "should not be based on TOEFL scores alone," the reality is that this test score carries tremendous weight and can make or break an application. Beyond admissions, however, the utility of the TOEFL score decreases, which Wongtrirat (2010) showed in a meta-analysis of whether TOEFL scores predict academic achievement. Granted, the extreme scores can help facilitate placement decisions, with the highest scores possibly exempting students from ESL courses and the lowest scores denoting the need for students to receive ESL instruction.

Still, a number of students want ESL instruction and see the benefit of it given that they are progressing toward a graduate degree in a language other than their native one. These students ought to have the opportunity to take ESL classes that have been shown to cater to their linguistic needs more than the mainstream classes.

This study has contributed to the wealth of research around the TOEFL test and its predictive validity. Additionally, the comparison of the TOEFL and graduate-level ESL placement exam, as well as the profile of the students who are taking the placement test, aids in future decision making about how to structure the ESL program to fit the needs of both the students and the university. Although there was only a slight trend indicating that TOEFL scores correlated with the placement-exam scores, the results can certainly facilitate advising with regard to the students with extreme scores. The students who score in the middle range can elect to enroll in the course (self-select)

or take the placement test to attempt to pass out of the requirement.

The results also suggest that the placement exam itself can be streamlined by taking out the essay section, given that the score plus summary scores can predict with near certainty the overall placement score. This, coupled with a smaller group of test takers, will save time, money, and energy with the administering of the test, grading, and advising.

The fact that the correlation was not high enough to disregard the placement exam in its entirety can be explained by the varying constructs of the two tests. The TOEFL is a massive-scale test of English designed to assess the proficiency of a wide range of test takers with diverse backgrounds and goals. The placement test, however, while still designed for students with a wide range of backgrounds, is more specialized in the sense that it is focused on the needs in a specific context, namely those of incoming international graduate students. Furthermore, the placement test matches the ESL curriculum so that deficiencies that are highlighted on the placement test are directly addressed in the subsequent ESL courses. In light of this system, the value of the local placement test holds strong.

Future Research

Ultimately, one of the central aims of this study is to better understand the linguistic needs of the international graduate student population at a northern California university, and the correlation analysis of test scores was the first step. The next step is to move beyond the initial stages that follow the students' arrival, and to track their progress throughout their respective graduate programs. Given that this is a diverse group of students with different needs and who engage in various academic disciplines, a qualitative or case study approach would create an opportunity to look deeper into the academic experience of international graduate students. Some questions to consider would be: Are these students well prepared after the ESL courses, or are they struggling? Should we continue to provide course work after the first year, and if so, what kind? Understanding the international graduate student population will allow the university to provide course work that is compatible with their needs.

Author

Daniel Moglen is a PhD student in the Department of Linguistics at the University of California, Davis, with interests in language assessment and second language acquisition. He holds a bachelor's degree in Linguistics from the University of California, Berkeley and a MA TESOL degree from the University of California, Davis.

References

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Burgess, T. C., & Greis, N. A. F. (1970). English language proficiency and academic achievement among students of English as a second language at the college level. Portland State University. Retrieved from http://files.eric.ed.gov/fulltext/ED074812.pdf

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL, 20,* 21-34.

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

Enright, M. K., & Quilan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing, 23*(3), 317-334.

Educational Testing Service. (2015). Retrieved from www.ets.org

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York, NY: Routledge.

Institute of International Education. (2015). 2014: A quick look at international students in the U.S. Retrieved from http://www.iie.org/Research-and-Publications/Open-Doors/Data/International-Students/Infographic

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

Kelleher, A. M. (2008). Placements and re-positionings: Tensions around CHL learning in a university Mandarin program. In A. W. He & Y. Xiao (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp. 239-258). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Krausz, J., Schiff, A., Schiff, J., & Van Hise, J. (2005). The impact of TOEFL scores on placement and performance of international students in the initial graduate accounting class. *Accounting Education, 14*(1), 103-111.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183-189.

Liu, O. L. (2011): Do major field of study and cultural familiarity affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education, 24*(3), 235-255.

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

Ng, J. N. K. (2007). *Test of English as a foreign language (TOEFL): Good indicator for student success at community colleges?* (Doctoral dissertation). Oregon State University, Corvallis.

Papajohn, D. (2006). Standard setting for next generation TEOFL Academic Test (TAST): Reflections on the ETS Panel of International Teaching Assistant Developers. *TESL-EJ, 10*(1). Retrieved from http://www.tesl-ej.org/wordpress/issues/volume10/ej37/ej37a1/

Raimes, A. (1987). Language proficiency, writing ability, and composing strategies: A study of ESL college student writers. *Language Learning, 37*(3), 439-468.

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 24*(3), 427-442.

Savignon, S. J. (1991). Communicative language teaching: State of the art. *TESOL Quarterly, 25*(2), 261-277.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*(4), 657-677.

Spolsky, B. (2007). *On second thoughts. Language testing reconsidered.* Ottawa, Canada: University of Ottawa Press.

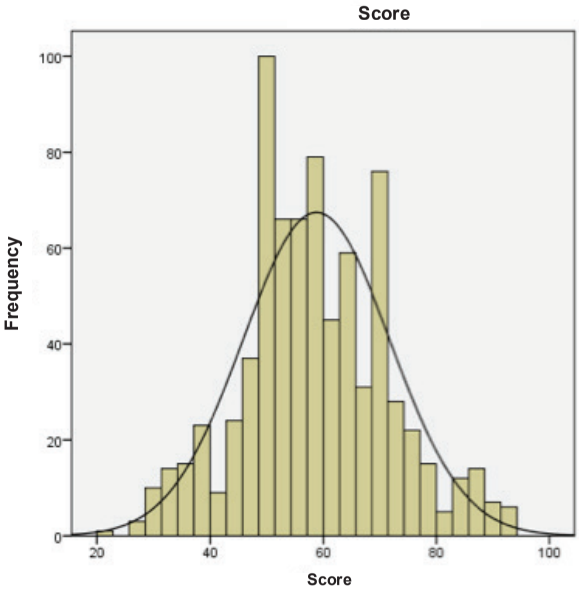
Wongtrirat, R. (2010). *English language proficiency and academic achievement of international students: A meta-analysis* (Doctoral dissertation). Old Dominion University, Norfolk, VA. Available from ProQuest Dissertations and Theses Database. (UMI No. 3417016)

Xi, X. (2010). How do we go about investigating fairness? *Language Testing, 27*(2), 147-170.

Zhang, H. (1996). *Academic achievement predicted by the test of English as a foreign language (TOEFL) across native language groups at Southern Connecticut State University* (Master's thesis). Southern Connecticut State University, New Haven.

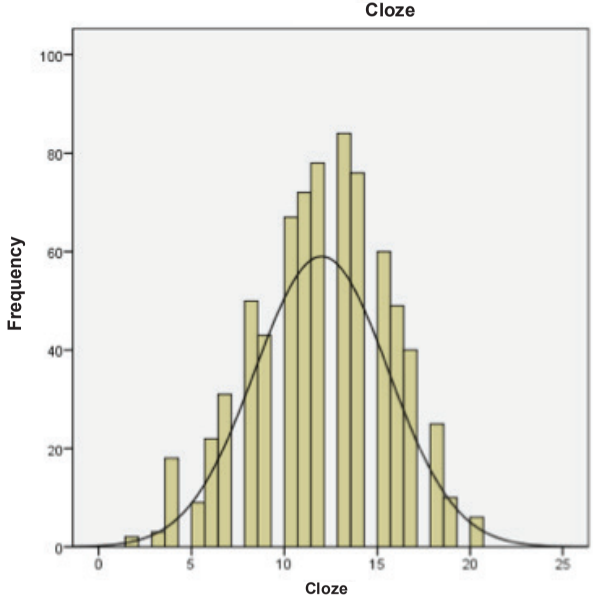
Appendix
Histograms/Scatterplot

Histogram: Placement Exam scores



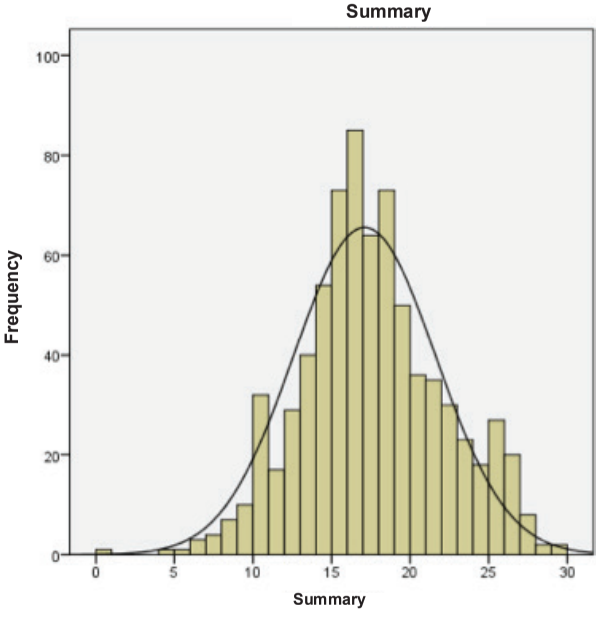
Mean = 58.8; *SD* = 13.0; *N* = 767

Histogram: Placement Cloze Test scores



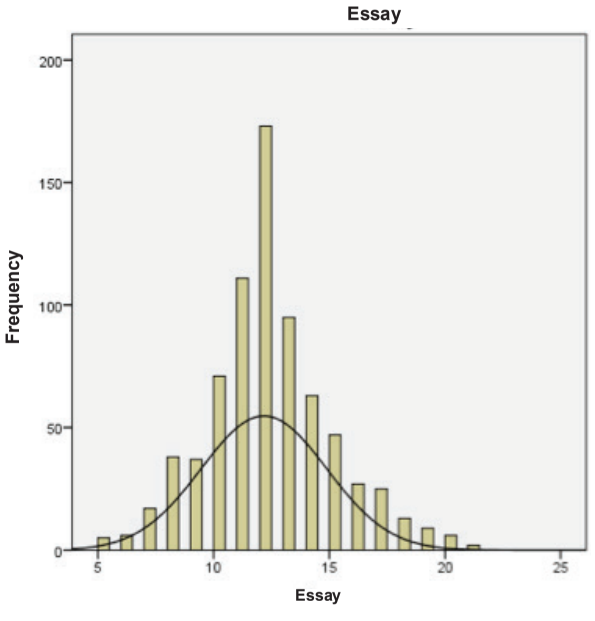
Mean = 12.0; *SD* = 3.6; *N* = 745

Histogram: Placement Summary scores



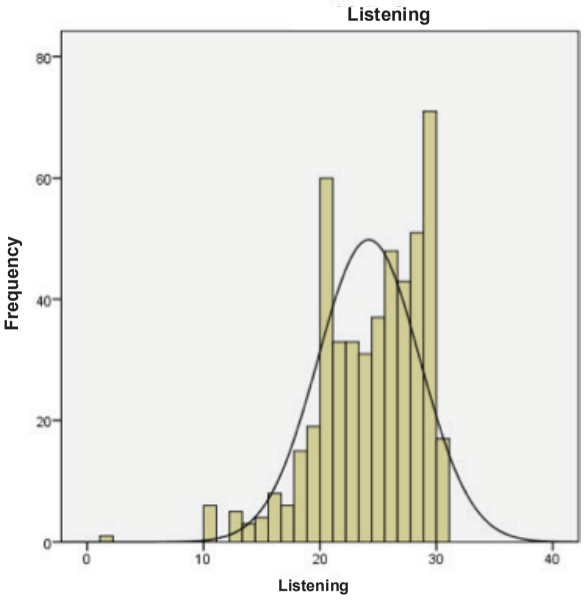
Mean = 17.1; *SD* = 4.5; *N* = 745

Histogram: Placement Essay scores



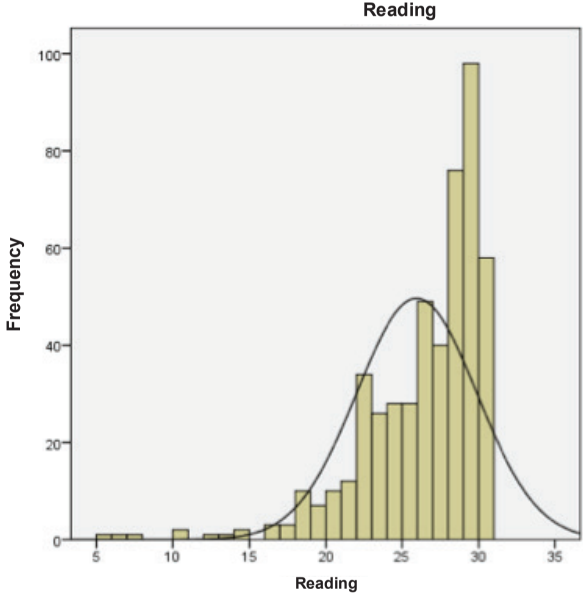
Mean = 12.2; *SD* = 2.7; *N* = 745

Histogram: TOEFL Listening scores



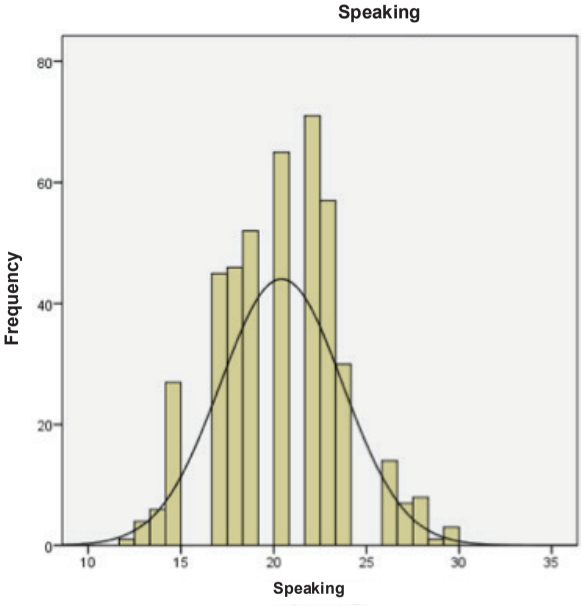
Mean = 24.2; *SD* = 4.4; *N* = 491

Histogram: TOEFL Reading scores



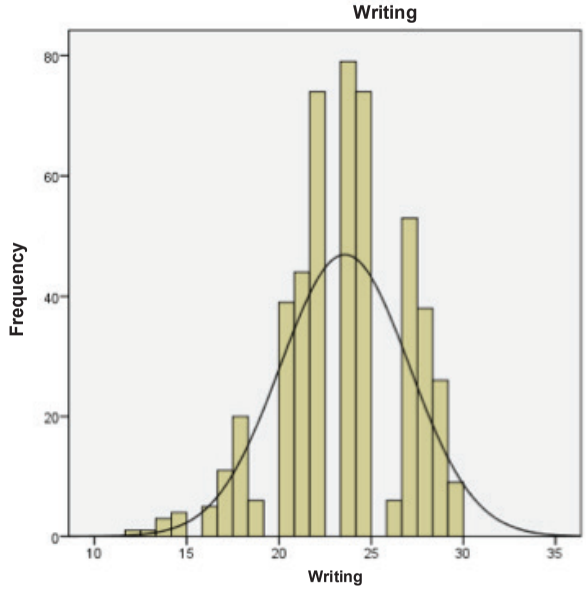
Mean = 25.9; *SD* = 3.9; *N* = 491

Histogram: TOEFL Speaking scores



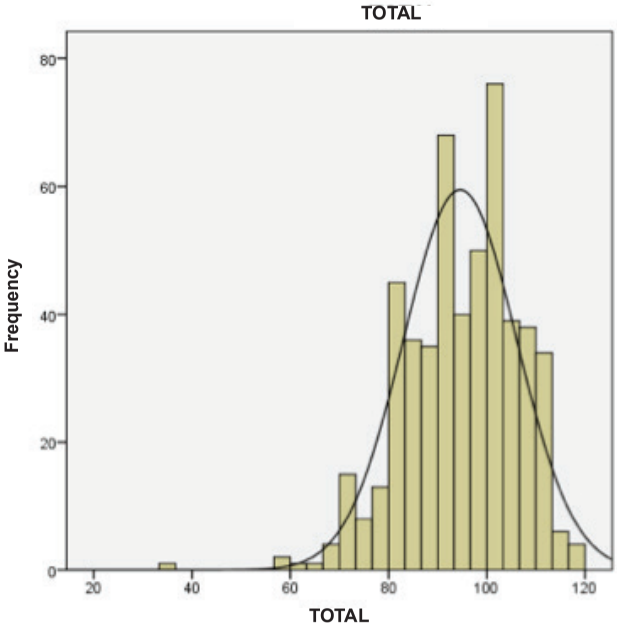
Mean = 20.4; *SD* = 3.3; *N* = 437

Histogram: TOEFL Writing scores



Mean = 23.6; *SD* = 3.5; *N* = 493

Histogram: TOEFL Total scores



Mean = 94.6; *SD* = 11.5; *N* = 516

Scatterplot: X-axis is TOEFL scores grouped (0-79, 80-84, 85-89, 90-94, 95-99, 100-104, 105-109, 110-114, 115-120), Y-axis is placement exam score