Test of English as a Foreign Language (TOEFL): Interpretation of multiple score reports for ESL placement

Kateryna Kokhan
Chih-Kai (Cary) Lin
University of Illinois at Urbana-Champaign

The vast majority of U.S. universities nowadays accept TOEFL iBT scores for admission and placement into ESL classes. A significant number of candidates choose to repeat the test hoping to get higher results. Due to the significant increase in the number of international students, the University of Illinois at Urbana-Champaign (UIUC) is currently seeking to find the most cost-effective ESL placement policy which would regulate the ESL placement of TOEFL repeaters. Since there is little published research examining students' multiple TOEFL iBT score reports, and there are no guidelines for the interpretation of multiple scores provided by the test publisher, this paper attempts to address the issue of interpretation and use of TOEFL iBT repeaters' scores for making ESL placement decisions in the context of UIUC. The main research question considered in our study was: Which TOEFL iBT scores (official highest, most recent, average or self-reported scores) are the best predictors of ESL placement? The findings indicate that the self-reported and the highest TOEFL iBT scores have the strongest association with the ESL placement results. The self-reported and the highest scores also demonstrate the highest classification efficiency in predicting ESL placement of TOEFL iBT repeaters. The results and implications of the study are discussed.

Key words: test repeaters, ESL placement, TOEFL, practicality, cost-efficiency

[1] Kate Kokhan, Department of Linguistics, University of Illinois at Urbana-Champaign, 4080 Foreign Languages Building, 707 S Mathews Avenue, Urbana, IL, USA, 61801. E-mail: kate.kokhan@gmail.

Introduction

The Test of English as a Foreign Language Internet-based Test (TOEFL iBT) measures the ability of non-native speakers of English to use and understand English in an academic context. According to the test publisher, there is no limit to the number of times test takers can take the test. Thus, many candidates who are not satisfied with their scores tend to repeat the test. Repeaters are examinees who have repeated the exam at least once, regardless of the time interval between the tests and/or the number of repeated tests (Yang, Bontya, & Moses, 2011). The main challenge that score users may experience when dealing with multiple score reports of TOEFL iBT repeaters is absence of any recommendations or guidelines from the test publisher regarding the interpretation of multiple scores for making admission and ESL placement decisions.

Score users who are considering making ESL placement decisions based on TOEFL iBT scores may ask themselves the following two questions:

1. To what extent can the TOEFL iBT score reports serve as an alternative to an ESL placement test?
2. To what extent can the TOEFL iBT score reports be indicative of the students' subsequent performance in the ESL classes?

Both of these questions deserve attention of researchers and stake holders; however, there are a number of extraneous factors completely unrelated to subsequent academic performance that are nowadays making score users pay more attention to the first question which directly concerns cost efficiency of test administration. Among such factors is a steady increase in the number of international students over the last decade. During the 2012-2013 academic year, the international student enrolment in the United States increased by seven per cent to 819,644 (Institute of International Education, 2013).

The vast majority of universities rely on such standardised tests as TOEFL and IELTS in making ESL placement decisions. Using the list of top 20 universities with the largest number of international students provided in the report, we reviewed the officially published TOEFL iBT score requirements for admission and the ESL placement policies in those educational institutions. Overall, the universities are very specific about the TOEFL iBT score requirements for admission; however, they differ significantly in terms of TOEFL iBT cut-off score[i] requirements for ESL placement: some campuses set higher cut-off score requirements (e.g., the Ohio State University (n.d.), TOEFL iBT 114; Indiana University – Bloomington (n.d.), TOEFL iBT 105), others set the scores

considerably lower (e.g., Michigan State University (n.d.), TOEFL iBT 80), and there are those that do not have any specific cut-off score policies at all (e.g., Columbia University (n.d.), York University (n.d.), etc.). The majority of the universities with the largest number of international students have local ESL placement tests and all of them offer ESL courses or their equivalents. However, we were not able to find any published guidelines for interpreting the scores of TOEFL iBT repeaters. The present study aims to provide ESL instructors, program directors and other score users, who are striving to find the most cost-efficient placement tool, with solid research evidence to refer to when interpreting multiple TOEFL iBT score reports of ESL students.

Previous research on test repeaters

There are several publications examining the interpretation of repeaters' scores (Boldt, Centra, & Courtney, 1986; Wightman, 1990; Yang, Bontya, & Moses, 2011; Zhao, Oppler, Dunleavy, & Kroopnick, 2010). Although most of these studies were conducted to explore multiple test scores from tests other than TOEFL and had different backgrounds and aims for interpreting scores of test repeaters, some of the findings are relevant to the present research and will be discussed below.

We discovered three major directions in research on the interpretation of multiple score reports. The first direction is the closest to the purpose of the present study and is focused on determining how multiple scores on one test can predict performance on another test. Zhao, Oppler, Dunleavy and Kroopnick (2010) investigated the validity of four approaches (average, most recent, highest-within-administration,[ii] and highest-across-administration[iii]) to using repeaters' Medical College Admission Test (MCAT) scores to predict Step 1 of the United States Medical Licensing Examination (USMLE) scores. The findings suggested that the best approach for computing repeaters' score for admission purposes was to take the average across all administrations. The authors pointed out that: "Students with the same MCAT average score are expected to perform the same on Step 1 exam regardless of the number of attempts to achieve that average. As such, admissions committees can have confidence that MCAT total scores computed by averaging across all administrations are comparable regardless of the number of times a student took the MCAT exam" (Zhao et al., 2010, p. S67). The authors warned against using the most recent, highest-within-administration and highest-across-administration score approaches because they were associated with a higher prediction error. These approaches can result in a so-called 'overprediction' of repeaters' Step 1 total scores as repeaters were expected to perform worse on

Step 1 than non-repeaters and repeaters who were tested fewer times but had the same MCAT total score computed using one of these approaches.

The purpose of the second direction is to examine the stability of scores across multiple testing occasions. For instance, Yang, Bontya, and Moses (2011) investigated repeater effects on score equating. They analysed a significant number of self-reported test scores across a wide range of administrations from a graduate admissions examination that was administered in a non-English language.[iv] The authors reported that repeater scores across testing occasions were fairly stable. However, there were large scale score gains/losses for a broad range of administrations. The test-retest correlation was 0.74 for Verbal and 0.72 for Quantitative section for the overall repeater group. The results also suggested that high-performing test takers might not improve their scores by retesting as would the low-performing examinees. The researchers admitted that further research would be needed in order to better understand repeater performance patterns while taking into consideration the demographic information.

To our knowledge, an ETS research report by Zhang (2008) is the only published paper analysing the scores of TOEFL iBT repeaters. The author's assumption was that under normal circumstances, the scores of test takers may vary insignificantly on the condition that the tests were repeatedly taken within a very short period of time and no intensive training occurred during this time frame. The repeater sample (N=12,300) analysed in the study was a self-selected sample of candidates who took one TOEFL iBT test in a month after taking the other. The author did not mention though how she knew that no TOEFL preparation had occurred prior to the second test session.

Zhang (2008) found that there was a good correlation between individual sections of the TOEFL iBT (0.78 for reading, 0.77 for listening, 0.84 for speaking, and 0.77 for writing), and between the total scores of two tests (0.91). She concluded that small changes were observed in the test scores between the first and the second test of the repeaters. The distribution of score changes resembled a symmetrical bell-shaped distribution and the total scores only slightly increased (Δ=3.74).

Even though the paper by Zhang (2008) was not investigating the relationship between repeaters' test scores and their ESL placement results, there were some data reported by the author that drew our attention. The standard deviation of the average score change as reported in the paper was quite high (9.50). Most universities usually have a narrow range of scores (about 20 points) between the minimum for admission and the minimum for exemption from additional

on-campus testing. For example, the minimum TOEFL iBT total score for general graduate admission to the University of Iowa (n.d.) is 81 and the minimum TOEFL iBT score for graduate admission to the George Washington University (n.d.) is 80. Students who scored at least 100 on the TOEFL iBT are exempt from additional English language testing at both of these universities. A large uncertainty in the average score change of TOEFL iBT repeaters may result in the misplacement of ESL students. Unlike the repeaters in Zhang (2008), the TOEFL iBT repeater population considered in the present study represents a typical international student population in a large public university in the United States and is not limited to short-term repeaters only.

Finally, the purpose of the third direction is to find out how various scores of test repeaters predict their future academic performance. Boldt, Centra, and Courtney (1986) were interested in finding the best way to treat multiple SAT scores of individuals in college admission. They investigated five methods of treating multiple test scores: simple average, weighted average (a) giving all the scores, except the latest, the same weight, and (b) giving all the scores, except the highest, the same weight, latest and highest. The combinations of variables and the methods of treating multiple scores were evaluated in regression equations developed using one-time testers. The results showed that the simple arithmetical average was as valid as the weighted average. Overall, all treatments of multiple scores resulted in under-prediction of actual college grades; however, the highest scores were slightly better predictors. For reasons of validity and simplicity, the authors recommended using an average of scores because the arithmetic average had the highest correlation with college GPA. However, as far as the effects on prediction are concerned, the average turned out to be the worst.

Similarly to Boldt, Centra, and Courtney (1986), Wightman (1990) conducted a study to determine what Law School Admission Test (LSAT) scores (initial, latest or highest) could most accurately predict the subsequent performance of test takers in law schools. The author used a least-squares regression analysis of different LSAT scores of test repeaters on first-year average in law school. The researcher found that, in general, repeaters tended to get lower LSAT scores than one-time test takers regardless of whether initial, latest or highest scores were considered. Another finding was that repeaters and one-time test takers performed comparably in their undergraduate academic work; however, one-time test takers earned generally higher GPAs in law school than LSAT repeaters. The author concluded that the average LSAT score was the best predictor for the majority of law schools. This finding was later confirmed by Thornton, Stilwell, and Reese (2006) and Thornton Sweeney, Marcus, and Reese (2010), who examined the validity of the most recent, highest and average LSAT

scores in terms of predicting first-year law school grades. Wightman (1990) also pointed out that the second-best predictor was the initial test score. However, this finding contradicted the results of Pitcher (1977), who found that the initial score produced "the most deviant predicted means" and suggested that the use of initial test scores would be unfair to applicants. Wightman (1990) assumed that the differences in the findings between the results of Pitcher (1977) and her study may be due to intensive test preparation or coaching among most recent repeaters, which could have inflated the most recent scores of test repeaters, making them less predictive of law school performance.

All three directions in the research on test repeaters described above provide some valuable insights about various approaches to the categorisation of multiple score reports and offer useful insights about the research methodology; however, all of these studies produced very mixed results and, most importantly, none of them can offer any solutions to the issue of cost effectiveness of test administration which is the main topic of the present research.

The present study

This study is motivated by the current need to find the most reliable and cost-effective ESL placement policy at the University of Illinois at Urbana-Champaign due to a significant increase in the number of international students. The main research question addressed in the paper is: Which scores (official highest, most recent, average or self-reported scores) are the best predictors of ESL placement? Note that an average score from a test repeater in this study refers to the average across repeats. In order to find a comprehensive answer to this question, we broke it down into two sub-questions:

1. Is there a significant difference among the highest, most recent, average, and self-reported scores of TOEFL iBT repeaters?
2. What is the relationship between ESL writing placement levels and TOEFL iBT scores (highest, most recent, average, and self-reported)?

Method

Data

The data analysed in this paper were collected over the period of Fall 2006 - Fall 2011. Two types of data were available for the analysis: self-reported and official. The self-reported data were collected during the registration process for the EPT. The self-reported information includes TOEFL iBT scores of the EPT candidates, their expected degree, gender, native country, and individual university ID numbers (UINs). The official data were collected by the Office of Admissions and Records during the students' application process. The data set contains the information about students' TOEFL iBT results, their country of origin, major, and department. Sanitising and matching of the official data with students' UINs were conducted by the Division of Management Information (DMI) at UIUC.

Overall, the data set contains the TOEFL iBT scores of 3032 students wherein the number of TOEFL iBT repeaters is 474 (15.6%). The score reports of TOEFL iBT repeaters that were incomplete and the TOEFL iBT scores of the students who were placed into specifically developed ESL courses regardless of their placement level (students from MA in TESL program and from the Economics Department) were excluded from the analysis. Thus, the total number of TOEFL iBT repeaters analysed in this study was 396.

Instruments

The Test of English as a Foreign Language Internet-based Test (TOEFL iBT) assesses the ability of non-native speakers of English to use and understand English at the university level. It consists of four sections: Reading, Listening, Speaking, and Writing. The scaled scores of each section range from 0-30. A total score is formed by adding the scores of each section, and ranges from 0-120. The Reading section consists of 3-4 reading passages and questions about them. The Listening section consists of six passages followed by questions about them. The Speaking section contains two independent and four integrated speaking tasks. The Writing section consists of one independent essay task and one integrated essay task (for details, see http://www.ets.org/toefl/ibt/about).

The English Placement Test (EPT) consists of two parts: a written essay test and a pronunciation test conducted in a form of an oral interview. It is usually administered several days before the beginning of students' first semester at UIUC. Completion of the ESL requirement is one of the conditions for non-native speakers of English for graduation from the University. Based on the

results of the written and oral parts of the EPT, students are placed into the academic writing and English pronunciation courses, respectively. English pronunciation courses focus on the sounds of natural speech, rhythm, stress, intonation, and the use of English spelling rules to guide the pronunciation of newly encountered words. Pronunciation courses are offered to a very limited number of ESL students (less than 5% of all ESL candidates each year) who speak with very strong accents. Due to a very small sample size of ESL students required to take a pronunciation course at UIUC and absence of any pronunciation components in the structure of the Speaking section of the TOEFL iBT, we decided to focus only on students' written EPT results. Given that this study focuses only on the written EPT that requires examinees to incorporate information from a lecture and an article, we examined the relationship between the ESL placement levels and TOEFL iBT total scores, TOEFL iBT writing scores, and TOEFL iBT combined scores from the Writing, Listening, and Reading sections.

The written EPT test is a paper-and-pencil placement test specifically developed to assess students' ability to write an academic essay using a combination of skills (reading, listening, speaking and writing). On a test day, EPT test takers are asked to write an argumentative essay using the information from both a lecture delivered by a trained EPT proctor and a reading passage from a journal article. The structure of the written EPT is very close to the integrated writing assignment in the TOEFL iBT. The main difference though is the presence of a group discussion in which EPT test takers participate right after they finish the first draft of an essay. Essays are rated holistically by specially trained EPT raters. Based on the EPT results, students are placed into three levels of the ESL composition classes (Level 1, 2 and 3). Level 1 is the lowest ESL level and Level 3 is the highest. Placement into Level 3 corresponds to an obligatory composition course for international undergraduate students (equivalent to the Composition I requirement for domestic undergraduate students), or to an optional advanced composition course for international graduate students. Placement into Level 2 corresponds to a required ESL writing service course, while placement into Level 1 corresponds to a required two-course sequence in ESL service writing courses for both undergraduate and graduate students.

The EPT has a long history of validity research reflected in the master's theses, doctoral dissertations and publications of the graduate students from UIUC who helped in developing, exploring and improving this test. Once a year, the EPT administration organises rater trainings for new EPT raters. In order to avoid any rater effects acquired over time – so called "rater drift" (Hoskens & Wilson, 2001; Wilson & Case, 2000), experienced raters regularly participate in recalibration sessions. The inter-rater reliability on the EPT was reported by Lee

and Anderson (2007). The reliability between two raters grading an essay was quite high and ranged from 0.75 to 0.95. Jang (2010) reported similar reliability estimates for the EPT. The percentage of misplaced students based on instructors' evaluations is reported to be relatively small – 10.7% (Cho, 2001). In addition, the majority of test takers believe that the EPT adequately reflects their English academic writing ability in that source-based writing is required in most academic courses (Cho, 2001). Perhaps one of the most appealing content-relevant features is the tight resemblance between the EPT tasks (e.g., information gathering, multiple drafts, group discussion, etc.) and classroom activities typically reflected in the ESL curriculum on which the EPT is based. More information about the test can be found in the EPT Bulletin (http://www.linguistics.illinois.edu/students/placement/documents/EPTbulletin_Jan2011.pdf) and in Kokhan (2012, 2013).

Participants

The scores of 396 TOEFL iBT repeaters who took the EPT over the Fall 2006 – Fall 2011 were analysed using SAS ver. 9.3 and R software ver. 2.13.2. The majority of TOEFL iBT repeaters at UIUC took the test twice (81.06%) and a sizeable fraction of candidates took it three times (14.90%). Among the test repeaters, 46.21% were undergraduate students, 34.60% were students pursuing a Master's degree, 17.42% were doctoral students, and 1.77% were exchange (non-degree) students. The ratio of female to male was 44.44% to 55.56%. Test takers from China constituted more than half of the test repeaters population (53.54%). They were followed by students from Taiwan (12.63%), South Korea (10.61%), Indonesia (3.18%), Turkey (2.53%), Thailand (2.02%) and the remaining nationalities were under 2% each (see Appendix A).

In order to determine how representative TOEFL iBT repeaters are of the entire EPT test taker population, we examined the differences between repeaters and non-repeaters (test takers who took the TOEFL iBT only once) with respect to their gender, expected university degree, native country, and placement results. The proportions of male and female candidates are very similar in both groups: female repeaters (44.4%) versus female non-repeaters (45.1%) and male repeaters (55.6%) versus male non-repeaters (54.7%).The comparison of TOEFL iBT repeaters and non-repeaters according to the expected degree shows that non-degree students are least likely to repeat the test: the fraction of non-degree students in the repeater population is only 1.8% whereas their fraction in the non-repeaters group is 8.7%. The fractions of PhD and undergraduate (UG) students in the repeater group are somewhat higher than those in the non-repeater groups which means that they are more likely to repeat the test: UG – 46.2% versus 43.3% and PhD – 17.1% versus 12.0%; however, the fraction of

Master's students in the repeater group is lower than in the non-repeater group: 34.6% versus 36%.

More than half of the test repeaters in this study were placed into Level 2 which corresponds to an advanced academic ESL writing course. 23.23% of test repeaters were placed into Level 1 (an introductory academic writing course) and 20.96% were placed into Level 3 (exemption and an optional writing course for graduate students or a required course equivalent of the Composition I requirement for undergraduate students). It is noteworthy that the repeaters placed into the three levels differ by their average TOEFL iBT scores: the students placed to higher levels have generally higher average TOEFL iBT scores (see Appendix B). The comparison of the average TOEFL iBT scores of repeaters and non-repeaters (see Appendix C) suggests that the students with higher scores seem to be less likely to repeat the test.

Procedures

A within-subject ANOVA test was conducted to see whether the highest, most recent, average and self-reported TOEFL iBT scores differ from each other. A follow-up Tukey test was used to locate sources of pair-wise differences among the four types of TOEFL iBT scores of test repeaters

We also examined the placement trends of TOEFL iBT repeaters over the period of Fall 2006–Fall 2011. We first sorted out the total highest, most recent, average, and self-reported TOEFL iBT scores according to a bin size of 4. For each bin, we calculated percentages of placement into each ESL level. After that, we represented the values graphically. In addition, the empirical placement trends in relation to TOEFL iBT scores were compared with model-based placement trends to see whether the observed trends triangulate with the model-based approach. More specifically, we performed the cumulative logistic regression in analysing the relationship between ordinal responses and continuous predictors. In this study, it is used to model the probability of test takers being placed in the three ESL writing placement levels with respect to their TOEFL iBT scores. It should be noted that the association between the EPT results and TOEFL iBT scores would be masked by the limited range of ESL placement levels (i.e., three levels only) if the analyses were done in ordinary linear regression. Nonetheless, due to the ordinal nature of ESL placement levels, the cumulative logistic regression is appropriate for analysing the association between ESL placement and TOEFL iBT scores as long as adequate variability in TOEFL iBT scores is observed (see Appendix B) for each ESL placement level.

Results

Analysis of the differences among highest, most recent, average and self-reported scores of TOEFL iBT repeaters

For this part of the analysis, we first calculated the average of TOEFL iBT scores of test repeaters. As shown in Table 1, self-reported total TOEFL iBT scores have the highest average (92.3[5]), followed by the highest (92.19), most recent (90.72) and average (88.05) scores.

Table 1. Summary statistics of TOEFL iBT total scores

Score	N	Mean	SD
Highest	396	92.19	8.65
Most recent	396	90.72	9.49
Average	396	88.05	9.26
Self-reported	396	92.35	8.52

In order to explore the significance of the differences among all scores, we ran a within-subject ANOVA. The results indicate that there is a significant difference among the highest, most recent, average and self-reported TOEFL iBT total scores ($F = 220.57$, $df = 3, 1185$ and $p < .0001$), suggesting that at least one of the TOEFL iBT scores differs from the others.

A follow-up analysis using Tukey pair-wise method shows that the average TOEFL iBT total score is the main source of pair-wise differences among the four types of scores from test repeaters (see Table 2). The average TOEFL iBT total score is the lowest among the four types of scores. In addition, self-reported TOEFL iBT scores do not differ significantly from both the highest and most recent scores. The difference between the highest and most recent scores is also not significant. These pair-wise results seem to suggest that the test repeaters are inclined to report their TOEFL iBT total scores that are very close to either their highest or most recent scores. It also suggests that the test takers are very aware of their own scores and are strategising about which ones to submit. In an earlier study, Kokhan (2012) matched self-reported and official scores for 91.9% of students in a similar data set and found that less than 3% of students significantly exaggerated their TOEFL scores.

Table 2. Tukey pair-wise comparisons among four types of TOEFL iBT total scores

Score	Difference	Lower bound	Upper bound	Adj. p-value
Highest-Average	4.14	2.49	5.78	<.0001
Most recent-Average	2.67	1.03	4.31	.0002
Self-reported-Average	4.29	2.65	5.49	<.0001
Most recent-Highest	-1.47	-3.11	.18	.099
Self-reported-Highest	.16	-1.49	1.80	.99
Self-reported-Most recent	1.62	-.02	3.27	.054

Relationship between ESL writing placement levels and TOEFL iBT scores (highest, most recent, average and self-reported)

Figure 1 illustrates an ideal scenario of ESL placement. In order to be usable for placement purposes, TOEFL iBT scores should significantly differ between the placement levels. For example, all students with low TOEFL iBT scores (e.g., 1-80) should be placed in Level 1 (the lowest ESL level), all students with higher TOEFL iBT scores (e.g., 81-100) should be placed in Level 2, and all students with the highest TOEFL iBT scores (e.g., 101-120) should be placed in Level 3.

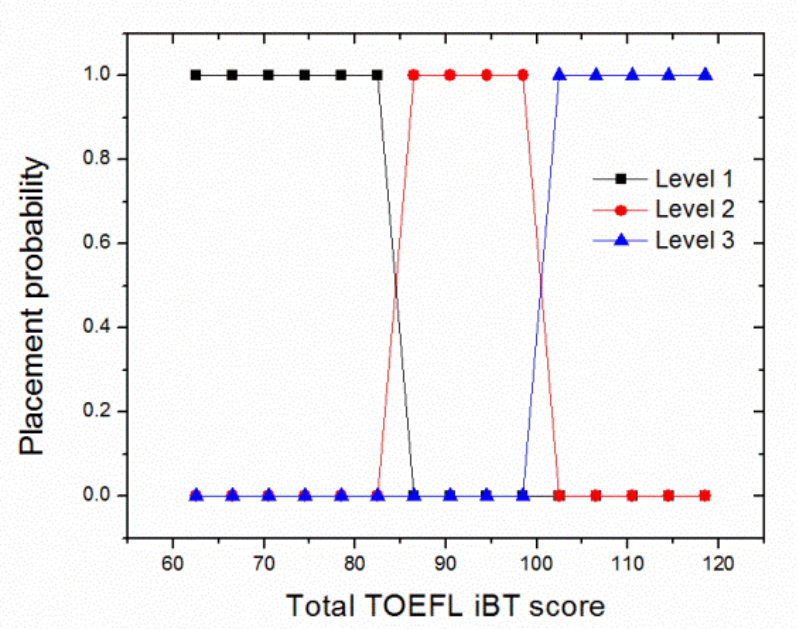

Figure 1. The ideal scenario of ESL placement based on the hypothetical TOEFL iBT score ranges for the three placement levels.

To check whether the actual ESL placement is close to the ideal scenario of placement, we built a series of graphs illustrating the placement trends in relation to the highest, most recent, average and self-reported total TOEFL iBT scores (see Figure 2). The results suggest that, no matter what scores we take (highest, most recent, average or self-reported), there is no distinct pattern of

placement which would help us to determine as to what scores of TOEFL iBT repeaters are better predictors of ESL placement. According to all four graphs, students with TOEFL iBT scores ranging from 70 to 105 have on average a 60% chance to be placed into Level 2 of the ESL writing classes. The repetition of the procedures with the highest, average, most recent and self-reported *section* scores yields very similar results, so they are omitted for reasons of space.

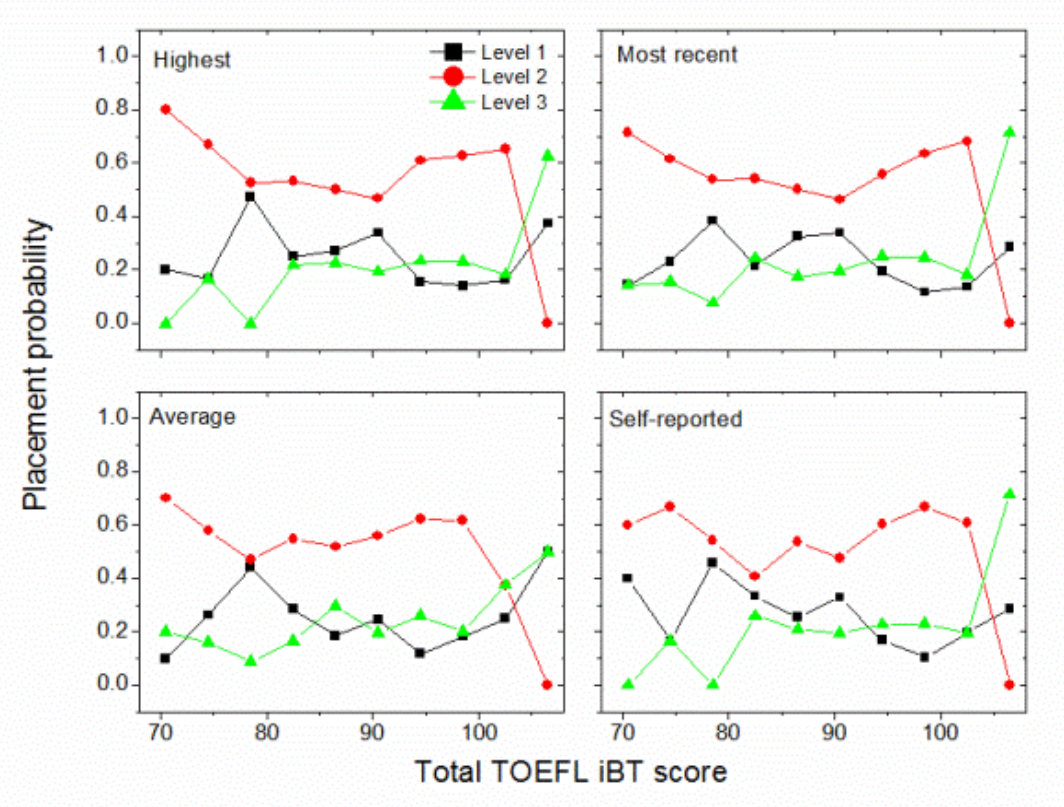

Figure 2. Probability of ESL placement in relation to the total highest, most recent, average and self-reported scores of TOEFL iBT repeaters.

To further explore the predictive capacity of the TOEFL iBT scores of test repeaters for ESL placement purposes, we conducted four cumulative logistic regression analyses based on the four types of TOEFL iBT total scores (highest, most recent, average and self-reported) respectively. The cumulative logistic model helps to collect two important pieces of information. First, it examines the extent to which the ESL placement levels are associated with test repeaters' TOEFL iBT scores. Second, the analysis can estimate the classification efficiency in predicting the ESL placement levels based on test takers' TOEFL iBT scores.

Inherent in the cumulative logistic regression model is the proportional odds assumption, according to which the logistic regression coefficient is the same for all cumulative logistic curves. In the present study, this implies that the

strength of association between TOEFL iBT scores and the odds of ESL placement is identical across all ESL placement levels (i.e. Level 1, 2 and 3). Table 3 shows the statistics of score test for the proportional odds assumption according to each TOEFL iBT score from test repeaters. This statistics can also be used as an alternative to goodness-of-fit statistics, such as deviance and Pearson chi-square.

Table 3. Score test for the proportional odds assumption

	Proportional Odds Assumption		
Score	Chi-square	DF	Pr > ChiSq
Highest	.28	1	.60
Most recent	.30	1	.59
Average	.38	1	.54
Self-reported	.07	1	.79

For each TOEFL iBT total score from test repeaters, the data appear to satisfy the assumption of proportional odds. For example, the cumulative regression analysis based on the highest TOEFL iBT scores shows that the proportional odds assumption is retained ($p = .60$), suggesting that there is no strong evidence of lack-of-fit. These results also indicate that the association between the TOEFL iBT total scores and ESL placement does not differ significantly at the higher and the lower ESL placement levels.

Table 4 shows an excerpt of logistic regression output from SAS ver. 9.3. The TOEFL iBT scores in this analysis are standardised according to the standard deviation of each score so that the results from the four cumulative logistic regression analyses are comparable.

Table 4. Summary table of cumulative logistic regression of ESL placement on four types of TOEFL iBT total scores respectively (n=396)

Score	Logistic regression coefficient	Standard error	Pr > ChiSq	Odds ratio estimates	Somers' D
Highest	.4115	.1011	<.0001	1.509	.180
Most recent	.3498	.0994	.0004	1.419	.171
Average	.3919	.1003	<.0001	1.480	.169
Self-reported	.4542	.1022	<.0001	1.575	.200

The logistic regression coefficient represents the degree to which ESL placement is associated with each TOEFL iBT score. Results show that for each TOEFL iBT total score, the logistic regression coefficient is significant at the .01 level, suggesting that there is a strongly significant association between the ESL

placement and each TOEFL iBT total score. For example, for one standard unit increase in the highest TOEFL iBT total score, the odds of a test taker being placed at a higher ESL level are 1.509 times (i.e., odds ratio estimate) the odds of the test taker being placed at a lower ESL level. Overall, the values of the estimated odds ratio indicate that the self-reported TOEFL iBT total scores from test repeaters have the strongest association with ESL placement results (1.575), followed by the highest (1.509), average (1.480) and most recent scores (1.419).

Somers' D is probably the most commonly used statistic in estimating the classification efficiency of cumulative logistic regression models (O'Connell, 2006). In this study, it is used as a rank order correlation between the predicted probability of ESL placement given the TOEFL iBT total scores and the observed ESL placement trends. Using different TOEFL iBT total scores in predicting the ESL placement levels, we found that self-reported TOEFL iBT total scores had the highest classification efficiency, followed by the highest, most recent and average total scores.

However, we must emphasise that there is a substantive difference between association and classification efficiency in an analysis such as this. The significant association between TOEFL iBT scores and ESL placement is to be interpreted in a probabilistic sense (i.e. a person with a higher TOEFL iBT score is more likely to be placed at a higher placement level) whereas classification efficiency is to be interpreted as the absolute predictive power of TOEFL iBT scores with respect to ESL placement results. Although the four types of TOEFL iBT scores are significantly associated with ESL placement, results show that the predictive power is low for any TOEFL iBT scores, ranging from .169 to .200. Furthermore, classification error rates based on the cumulative logistic model are above 57% regardless of which TOEFL iBT score is used.

Given that the current study focused on the EPT written test only and that the EPT required examinees to incorporate information from a lecture and an article, we also conducted separate analyses that looked into the relationship between the ESL placement levels and TOEFL iBT Writing scores only, and the relationship between the ESL placement levels and TOEFL iBT combined scores based on the Writing, Listening and Reading sections. Results based on the TOEFL iBT Writing scores showed that the odds ratio estimates ranged from 1.314 to 1.448 and the classification efficiency estimates were from 0.135 to 0.171; results based on the TOEFL iBT combined scores showed that the odds ratio estimates were between 1.236 and 1.470 whereas the Somers' D were between 0.103 and 0.160. Similarly, strongly significant association was observed between the ESL placement levels and TOEFL iBT scores; however, the classification efficiency was low for any TOEFL iBT scores from test

repeaters based on either the TOEFL iBT Writing section scores only or the combined scores for writing, listening and reading.

Discussion

We must be cautious in generalising our results to other educational institutions. However, we would like to emphasise at the beginning of this section that our study is based on a comprehensive campus-wide data set collected over a multi-year period and provides a detailed methodological approach to analysing the relationship between ordinal ESL placement levels and continuous standardised language test scores, from which other ESL placement programs could benefit.

The results of this study suggest that the average TOEFL iBT scores are significantly lower than the other three TOEFL iBT scores. In addition, the differences between the highest and most recent scores, between the highest and self-reported scores, and between the most recent and self-reported scores are not significant. In fact, the highest, most recent and self-reported TOEFL iBT scores from test repeaters at UIUC do not differ by more than 1.63 points. Thus, we can conclude that the test repeaters at UIUC tend to self-report either their highest or most recent TOEFL iBT scores.

The analysis of ESL placement trends for the period of Fall 2006–Fall 2011 in relation to the total highest, average, most recent and self-reported TOEFL iBT scores indicates that none of the scores are suitable for making accurate ESL placement decisions since, in some cases, students with very low TOEFL iBT scores can be placed into the highest ESL level and students with very high TOEFL iBT scores can be placed into the lowest ESL level. The results of the cumulative logistic regression analyses demonstrate that the absolute predictive power of all four types of TOEFL iBT scores is generally low. However, the self-reported TOEFL iBT total scores, followed by the highest TOEFL iBT scores, have the strongest association with the ESL placement results and they also demonstrate the highest classification efficiency in predicting ESL placement. We believe that the significant association between the TOEFL iBT scores of test repeaters and their ESL placement results could be explained by the fact that both tests deal with academic English language proficiency. The poor predictive capacity of TOEFL iBT scores may be due to the difference between the intended purposes of the TOEFL iBT and the EPT. TOEFL iBT scores are primarily used for admission purposes as a pre-arrival measure of students' ability to use and understand English in the academic environment. On the other hand, the EPT is a local placement test administered several days before

students' first semester at UIUC. It simulates a real classroom environment and is better aligned with the goals of the University's ESL writing curriculum. The low predictive power of TOEFL iBT scores may be due to its lack of sensitivity to the specific objectives embedded in the ESL curriculum, from which the test specification of the EPT is derived.

Even though the results of this research firmly indicate that none of the four types of scores considered in this paper are reliable for ESL placement, the findings of this study prompt us to suggest that the self-reported or the highest TOEFL iBT scores might be the best option for considering when making ESL placement decisions under the circumstances with limited financial resources and in the situation where it is impractical to require all students to take an ESL placement test. The collection of self-reported TOEFL iBT scores from students upon their arrival on campus, as it is currently done at the University of Illinois, requires much less financial and human resources compared to the administration of English as a Second Language placement tests. The collection of students' highest TOEFL iBT scores may cost virtually nothing since TOEFL score reports are sent directly to score users during the application process and can be requested by program directors from local Offices of Admissions and Records. Making early ESL placement decisions based on students' TOEFL iBT scores would allow program administrators to be more efficient in planning and scheduling ESL courses.

The review of the literature performed for this study prompted the question which is beyond the scope of the present study but can be addressed in future research: "To what extent can the TOEFL iBT score reports be indicative of the students' subsequent performance in the ESL classes?" It would be interesting to explore the link between students' test results and their actual performance, achievement or grades in the ESL classes. However, in the context of those universities like UIUC where ESL classes are graded on the 'satisfactory/unsatisfactory' basis, this question may be harder to answer.

In conclusion, we found that TOEFL iBT scores may not be a reliable predictor of ESL placement. However, if placement test administration is not possible or practical for any reason, the highest or self-reported TOEFL iBT scores may be the most cost-efficient option for ESL placement of students who took TOEFL iBT more than once.

Acknowledgements

We thank Professor Fred Davidson from the University of Illinois at Urbana-Champaign for his help with getting the official TOEFL data. We are also thankful to two anonymous reviewers and the editors of PLTA for their helpful comments and suggestions.

References

Boldt, R. F, Centra, J. A., & Courtney, R. G. (1986). *The validity of various methods of treating multiple SAT scores*. (College Board Report No. 86-4). New York: College Entrance examination Board.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Cho, Y. (2001). *Examining a process oriented writing assessment in a large-scale ESL testing context*. Unpublished Doctoral Dissertation. University of Illinois, Urbana-Champaign.

Columbia University. (n.d.). *Admission overview for international students*. Retrieved from http://www.columbia.edu/cu/isso/admit/.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*, 121-145.

Indiana University Bloomington. (n.d.). Transfer admission standards. Retrieved from http://www.indiana.edu/~iuadmit/apply/transfer/standards.shtml.

Institute of International Education. (November 11, 2013). *Open Doors 2013: International Students in the United States and Study Abroad by American Students are at All-Time*. Retrieved from http://www.iie.org/Who-We-Are/News-and-Events/Press-Center/Press-Releases/2013/2013-11-11-Open-Doors-Data#.Uo-mA8Ssh8F.

Jang, S. Y. (2010). *The development and evaluation of a systematic training program for increasing both rater reliability and rating accuracy* (Doctoral dissertation). Retrieved from http://hdl.handle.net/2142/15599.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision making: Placement trends and effect of time lag. *Language Testing*, *29*(2), 286-303.

Kokhan, K. (2013). An argument against using standardised test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing*, *30*(4), 467-489.

Lee, H.-K. & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), 307-330.

Michigan State University. (n.d.). International student information. Retrieved from http://grad.msu.edu/apply/docs/EnglishLanguageRequirements.pdf.

New York University. (n.d.). *International students*. Retrieved from http://www.nyu.edu/admissions/undergraduate-admissions/apply/freshmen-applicants/international-students.html.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables. Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage Publications.

Pitcher, B. (1977). *The validity of Law School Admission Test scores for repeaters*. (Report No. LSAC-775). Princeton, NJ: Law School Admission Council.

The George Washington University. (n.d.). *Graduate Admissions: English Language Requirements*. Retrieved from http://graduate.admissions.gwu.edu/english-language-requirements.

The University of Iowa. (n.d.). *Graduate Admissions: English Proficiency Requirements for International Applicants*. Retrieved from http://grad.admissions.uiowa.edu/graduate-programs/english-requirements-admission-graduate-college.

The Ohio State University. (n.d.). *Graduate admissions: Frequently asked questions*. Retrieved from http://gradadmissions.osu.edu/faqs/FAQSubject.aspx?sub=Requirements/Deadlines&cat=%201&tl=before%20you%20apply.

Thornton, A. E., Stilwell, L. A., & Reese, L. M. (2006). *The validity of law school admission test scores for repeaters: 2001 through 2004 entering law school classes*. (LSAT Technical Report 06-02). Newtown, PA: Law School Admission Council.

Thornton Sweeney, A., Marcus, L. A., & Reese, L.M. (2010). *The validity of Law School Admission Test scores for repeat test takers: 2005 through 2008 entering Law School classes*. (LSAT Technical Report 10-02). Newtown, PA: Law School Admission Council.

Wightman, L. F. (1990). *The validity of law school admission test scores for repeaters: A replication*. (LSAC Research. Report No. 90-02). Newtown, PA: Law School Admission Council.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. V, pp. 113-133). Stamford, CT: Ablex.

Yang, W. L., Bontya, A. M., & Moses, T. P. (2011). *Repeater effects on score equating for a graduate admissions exam*. (ETS Research Report No. RR-11-17). Princeton, NJ: ETS.

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT*. (ETS Research Memorandum No. RM. 08-05). Princeton, NJ: ETS.

Zhao, X., Oppler, S., Dunleavy D., & Kroopnick, M. (2010). Validity of four approaches of using repeaters' MCAT scores in medical school admissions to predict USMLE Step 1 total scores. *Academic Medicine*, *85*, S64–S67.

Appendix A

Distribution of TOEFL iBT repeaters and non-repeaters according to the native country

Native country	Total population		Non-repeaters		Repeaters	
	Frequency	%	Frequency	%	Frequency	%
China	1464	49.6	1252	48.9	212	53.54
South Korea	387	13.1	345	13.5	42	10.61
Taiwan	379	12.8	329	12.9	50	12.63
India	89	3.0	86	3.4	3	0.76
Thailand	51	1.7	43	1.7	8	2.02
Turkey	58	2.0	48	1.9	10	2.53
Japan	42	1.4	39	1.5	3	0.76
Malaysia	37	1.3	32	1.3	5	1.26
France	35	1.2	32	1.3	3	0.76
Brazil	29	1.0	28	1.1	1	0.25
Indonesia	29	1.0	16	0.6	13	3.28
Iran	24	0.8	19	0.7	5	1.26
Colombia	20	0.7	15	0.6	5	1.26
Kazakhstan	19	0.6	14	0.5	5	1.26
Other countries	291	9.8	260	10.1	31	7.82
Total	2954	100	2558	100	396	100

Appendix B

Distribution of TOEFL iBT repeaters according to the EPT placement levels

EPT placement	Frequency	Percent	Mean iBT total score	Std Dev
Level 1	92	23.23	84.83	11.24
Level 2	221	55.81	88.58	8.60
Level 3	83	20.96	90.22	7.53
Total	396	100	88.05	9.26

Appendix C

Distribution of TOEFL iBT non-repeaters according to the EPT placement levels

EPT placement	Frequency	Percent	Mean iBT total score	Std Dev
Level 1	492	21.3	90.42	11.62
Level 2	1226	53.1	92.32	8.48
Level 3	589	25.5	94.62	8.02
Total	2307	100	92.50	9.24

Note. Students from MA in TESL program and students majoring in Business and Economics (N=251) were excluded from this analysis since they are required to be enrolled into the ESL writing courses specifically developed to meet their language needs.

Endnotes

[i] A cut-off score (or a cut-point) is "that score at or above which students will be classified one way and below which students will be classified differently" (Brown, 1996, p. 249). In the case of ESL placement of international students, such a cut-point separates students who need some remedial ESL courses from those whose language skills satisfy the academic requirements set by the university.

[ii] Highest-within-administration score is a score from the administration in which students received the highest total score (Zhao et al., 2010).

[iii] Highest-across-administration score is a total score computed by summing the highest section scores across administrations (Zhao et al., 2010).

[iv] Yang, Bontya and Moses (2011) did not explicitly mention the name of the test.