Examining Results and Establishing Benchmark Data from the TOEFL ITP Test

Rosalind Warfield, Rachel Laribee, Roger W. Geyer
The American University of Iraq, Sulaimani (AUIS)
Sulaimani – Kirkuk Main Road, Raparin
Sulaimani Governorate
Kurdistan Region, Iraq
rosalind.warfield@auis.edu.iq, Rachel.laribee@auis.edu.iq, roger.geyer@auis.edu.iq

Abstract. This exploratory research study reports on TOEFL ITP test scores of students at the American University of Iraq, Sulaimani (AUIS). This test replaces a "homegrown" placement entrance exam in use at AUIS for the last five years. Two hundred seventy-one scores were recorded for the test given February 21, 2013, which represents 85% of students enrolled in the Academic Preparation Program (APP). Variables of interest include students' APP level, native language, gender and their relationships to test scores and students' grades. Descriptive statistics were collected on 85 females and 187 males across four levels of APP. Kurdish is the predominant language (73.16%) followed by Arabic (21.32%); the remaining students speak a variety of different languages (5.52%). Mean differences in scores across APP levels were statistically significant across all levels except 3 and 4. Examining test scores and their grade equivalencies across reading and writing/grammar ITP scores showed weak correlations ($r = .156$ and $r = .109$, respectively) but a moderate correlation ($r = .334$) for listening. Research is ongoing.

Keywords: ITP, Iraq, TOEFL, AUIS, Kurdish, EFL

1 INTRODUCTION

The American University of Iraq, Sulaimani, in the northern Kurdish region, began operations in 2007. The Academic Preparatory Program (APP), a four-semester pre-academic course of English language training, has used an English language test written by instructors to place students in appropriate levels. Although this test was never psychometrically evaluated, the general consensus of the APP administration was that the test was reliable and valid based on student performance in the assigned levels. However, the availability of the TOEFL ITP, a professionally designed and universally recognized test from the Education Testing Service (ETS) ("TOEFL ITP Test," 2013) in Princeton, NJ, seemed an opportunity to make our evaluation and placement process a more accurate indicator of student ability (and perhaps of future success), and the decision was made to use the ITP exclusively for placement beginning February, 2013. The current study is an attempt to begin to validate that assumption.

2 THE STUDY

The goal of this study is to establish a benchmark of data from which appropriate research questions can be raised and answered. These questions are driven by the desire to understand our student population at AUIS in order to improve and refine student placement and to better gauge student readiness for advancement through the APP and into the undergraduate program. Further, it is also of considerable interest to examine how gender, ethnicity, and student background (including socio-economic standing) affect student learning at the APP and academic levels at AUIS.

2.1 Purpose

The purpose of this preliminary study is to analyze results of the first data generated from ITP testing. Since such a large population of students was tested, it is especially important to document and capture all data relating to this event. Through this initial effort we hope to articulate appropriate research questions that will lead to a better understanding of how to prepare Iraqi students for success in an American style university setting.

2.2 Participants

Of 320 students currently enrolled in the APP program, 276 students took the exam. Twenty-one students took the exam the following week, and that data is not included in this study. The remaining 23 students were either absent or tardy for their scheduled exam, resulting in their exclusion from this study. Of the 276 exam scores reported, five were discarded for malformed or incomplete data.

2.3 Instruments

The TOEFL ITP test is part of the ETS Global assessment series which measures students' English language skills in three areas: 1) Listening Comprehension, 2) Structure and Written Expression, and 3) Reading Comprehension ("Scores Overview," 2013). The paper-based test uses 100 percent academic content to gauge the abilities and proficiency of nonnative English speakers, according to the ETS website. Test content is available in two levels. Level 1 is a two-hour test designed for high intermediate to advanced students. Level 2 is shorter, one hour and 10 minutes, and is designed for high beginning to intermediate-level students.

2.4 Procedures

The Level 1 test was given to all students on February 21, 2013. Scoring sheets were sent to ETS in Baghdad the next day. Individual scores were received for the three areas outlined above on March 3, 2013, and imported into an Excel spreadsheet where additional information was added including students' principal language, gender and current APP level. The final APP grades of 213 students from the fall, 2012 semester (individual grades for grammar, reading, listening and writing) were added to the spreadsheet. New spring, 2013, students and students who withdrew before completing the fall, 2012, semester do not have final grades included in this data, which explains the difference between the 271 ITP test-takers and the 213 students who passed in the previous APP level. We presumed that the ITP subtests and the APP grammar, reading, and listening areas of concentration had basic equivalencies. After this data was compiled and double-checked, it was imported into IBM's SPSS statistical software, v. 20, and the following statistical results were generated.

3 RESULTS

3.1 Descriptive Statistics

Test results by gender are presented in Table 1. ETS maps the Level 1 TOEFL ITP test scores to the score levels of the Common European Framework of Reference for Languages (2013) for better clarification and interpretation.

Table 1. Descriptive Statistics: Gender

Gender		Listening	Grammar	Reading	Total_ITP_Score*
Males, N = 186	Mean	44.89	41.33	40.03	420.82
	SD	5.637	5.399	6.472	48.042
Females, N = 84	Mean	44.23	40.90	39.61	415.77
	SD	6.444	5.053	6.534	51.033
Total, N = 270	Mean	44.68	41.20	39.90	419.25
	SD	5.896	5.288	6.482	48.953

* Subtest scores are mapped to the Common European Framework of Reference for Languages

Although males scored higher, the comparative mean differences were not statistically significant for subtests and total scores.

Table 2 presents test score results broken down by APP level. The box plot in Figure 1 visually shows the differences between levels.

Table 2. Descriptive Statistics: APP Level

APP_Level, N		Listening	Grammar	Reading	Total_ITP_Score
1, N = 54	Mean	39.00	36.80	35.61	371.39
	SD	4.400	3.693	4.016	23.493
2, N = 112	Mean	43.43	39.82	38.72	406.55
	SD	4.427	3.790	5.413	32.291
3, N = 69	Mean	49.54	44.80	43.23	458.52
	SD	3.837	4.060	6.746	35.999
4, N = 36	Mean	47.53	45.03	43.53	453.61
	SD	6.336	6.421	7.004	59.605
Total, N = 271	Mean	44.65	41.18	39.89	419.03
	SD	5.914	5.292	6.471	48.994

\

Figure. 1. Box plots of Total Scores, N = 271
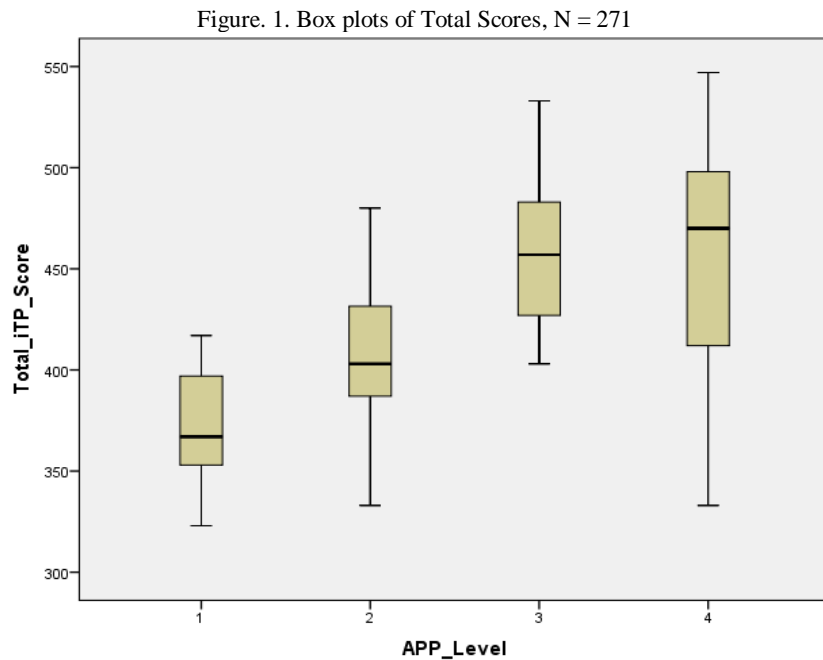
Tukey's test revealed significant differences ($p < .001$) between all levels except for Levels 3 and 4 where the mean was actually lower for Level 4. This result is interpreted in the Discussion section below.

Table 3 notes the results given the principal language of the participants.

Table 3. Descriptive Statistics: Language

Language		Listening	Grammar	Reading	Total_ITP_Score
Kurdish, N = 199	Mean	44.38	41.27	39.77	418.02
	SD	6.005	5.503	6.320	49.660
Arabic, N = 57	Mean	46.49	41.23	40.91	428.81
	SD	5.138	5.043	7.115	47.742
Other, N = 15	Mean	41.20	39.80	37.60	395.27
	SD	5.609	2.808	5.501	35.919
Total, N = 271	Mean	44.65	41.18	39.89	419.03
	SD	5.914	5.292	6.471	48.994

The relationships between the APP grades and ITP test scores were analyzed using Pearson's correlation coefficient. The results are reported in Table 4.

Table 4. Correlations: Subtest scores and APP equivalents, N = 271

Subject	Correlation Coefficient (r)
Grammar	.156
Reading	.109
Listening	.334
Total Score	.221

4.2 Feedback from test takers

Students in APP Levels 2, 3, and 4 said that they found the exam to be challenging yet appropriate for their English ability. Though this attitude pre- or post-exam did not always correlate to high marks on the ITP, all students in those levels said that the exam was fair and passable. Students in our lower level, APP 1, thought the listening portion of the exam was extremely difficult.

4.2.1 Results from the TOEFL

Students trust TOEFL scores more than the APP placement exam. In the past, students argued that our exam placement score was not accurate. Though the results from the TOEFL generally agreed with our exam (all but five placements from our homemade placement exam were the same as the TOEFL placements), students accepted the TOEFL placement without complaint.

4.3 Evidence from test administration – Lessons Learned

Administering an exam to 276 students at one time is a challenge. We recommend that the proctors walk through the exam and thoroughly understand how to administer it before students arrive. Some proctors and students had trouble understanding how to fill out the answer sheet. Students in Levels 1 and 2 need extra time to fill out the answer sheet.

Assessing the data was much easier than evaluating results of our placement exam. We received the results from ETS within a few days.

5 DISCUSSION

The following is a discussion of results presented above.

5.1 Gender, Language and APP Levels

Although results seem to indicate that students are correctly placed in Levels 1 through 3, Level 4 scores actually dropped compared to Level 3, M = 453.61 vs. M = 458.52, respectively. This surprising result was very concerning but explainable once we looked beyond the numbers. A large number of students were repeating Level 4 after failing the previous semester. Examining a histogram of Level 4 results (Figure 2) indicates a negative skew and seems to represent two distinct groups. Looking at the academic history of these students is revealing. The majority of Level 4 students who took this exam had tested into Level 4 in the fall. They did not have the benefit and practice of previous APP classes in test-

taking skills. Additionally, there was no personal incentive for them to take the test seriously because, unlike students in the lower levels, they did not have the opportunity to move up a level based on the results, a "carrot" we offered to all students as a motive to do their best.

Figure. 2. Histogram of Level 4 Scores, N = 36
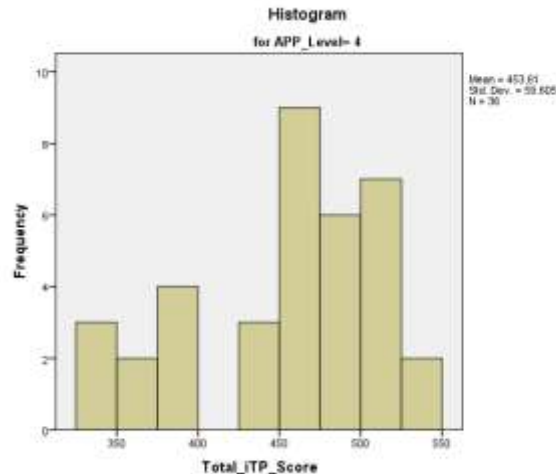
5.2 – Relationship between Scores and Grades

The relationship between ITP subtest scores and APP grades was not strong, although a moderate correlation between Listening, r = .334 is noted. The overall correlation between grades and test scores was not compelling with r = .221.

Success in the undergraduate programs and the Academic Preparatory program depends not only on a student's English language ability but also on test-taking strategies and practice with standardized exams. Therefore, passing an Academic Preparatory class based on more than just one exam (i.e. homework, quizzes, essays, multiple-section exams and a final comprehensive exam) is a better indicator of success in Academic classes than are the results of a single standardized test.

6 CONCLUSION

A TOEFL exam has many benefits. From the administrative side, having to create, score and assess an exam for prospective students takes a huge toll in time and energy. Using the TOEFL ITP eliminates those costs.

Using a TOEFL exam precludes students from arguing with their level placement. Because a TOEFL exam is internationally recognized and respected, students seem to value and accept their TOEFL results more than they have done our "homegrown" exam.

APP has decided to use the TOEFL ITP as our placement exam beginning in fall, 2013. Though we considered our previous instrument to be valid and reliable, the cost in time and energy was too high. With a growing program testing hundreds of students each semester, we believe that moving to the TOEFL ITP is cost effective and labor-saving.

6.1 Ongoing Research

All students took the ITP Level 1 (intermediate to advanced) test. Should the ITP Level 2 (high beginning) test be given to our Level 1 APP students? How will these students perform academically as they move through higher APP levels and into the academic program? Will performance on the TOEFL ITP exam provide any kind of reliable predictability of future high-stakes testing results and future academic performance? Are there are other factors that influence test results? These are some of the questions and issues that will be considered as we collect and analyze more data.

Acknowledgments

References

The TOEFL ITP Tests. (2013). Retrieved March 1, 2013, from
http://www.etsglobal.org/Tests-Preparation/The-TOEFL-Family-of-Assessments/The-TOEFL-ITP-Tests

Scores Overview. (2013). Retrieved March 1, 2013, from http://www.etsglobal.org/Tests-Preparation/The-TOEFL-Family-of-Assessments/The-TOEFL-ITP-Tests/Scores-Overview

The Common European Framework of Reference for Languages. (2013). Retrieved March 1, 2013, from http://www.etsglobal.org/Research/CEFRL

Rosalind Warfield has been Director of the Academic Preparatory Program at the American University of Iraq since 2008. She is a graduate of Randolph-Macon Woman's College and Boston College. She previously taught writing and humanities at Hampden-Sydney College and the University of California at Santa Cruz. Her publications include short stories and "Clear Writing Means Clear Thinking: Two Narratives of the Collaborative Teaching of James Madison's *Federalist 10 & 51,"* with James Pontuso.

Rachel Laribee has been Deputy Director of the Academic Preparatory Program at the American University of Iraq since 2010. She graduated from the American University with a MA in International Development and from Saint Mary's College of Maryland with a BA in Chinese Studies and Mandarin. She has concluded two field research publications, "The China Shop Phenomenon: Trade Supply within the Chinese Diaspora in South Africa" published in the *Afrika Specturem*, and the Tibetan Sky Burial" published in the *River Gazette*.

Roger W. Geyer is an associate professor at the American University of Iraq, Sulaimani where he teaches undergraduate and graduate courses in Information Technology. He received his BA degree in economics from New York University and an MS degree in mathematics instruction from Old Dominion University, Norfolk, Virginia. He received his PhD degree in instructional technology from the Curry School, University of Virginia, Charlottesville, Virginia. He has an extensive background in business and information systems along with numerous certifications including Microsoft's MCSE. He is the author of many presentations and papers regarding the use of technology in education and business. His dissertation was published as a book, *Internet-Savvy Youth: What the kids have to tell us*

about using the Internet. He has a wide range of research interests including knowledge acquisition, storage and distribution across the Internet, and collective intelligence systems.