ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-83
ETS RR–18-43

An Investigation of the Predictive Validity of the *TOEFL iBT*® Test at an English-Medium University in Turkey

John O'Dwyer

Elif Kantarcıoğlu

Carole Thomas

December 2018

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

An Investigation of the Predictive Validity of the *TOEFL iBT*® Test at an English-Medium University in Turkey

John O'Dwyer, Elif Kantarcıoğlu, & Carole Thomas

Bilkent University, Ankara, Turkey

This study reports on an investigation of the predictive validity of the *TOEFL iBT*® test in an English-medium institution (EMI) in a non-target-language context, namely, Turkey. The relationship between TOEFL iBT scores and academic performance was explored in a cohort of 286 undergraduate students, as was the TOEFL iBT's relationship with an institutional English proficiency exam (Certificate of Proficiency in English [COPE]) used as a benchmark for faculty entry. Performance measures included scores on TOEFL iBT and COPE, grade point averages (GPAs) for content and English for academic purposes (EAP) courses over 2 freshman semesters, and freshman language instructor evaluations of students' freshman EAP performance. Correlations between test scores confirmed a moderate to moderately high predictive validity for content course GPAs and English-language course GPAs, respectively, and for the TOEFL iBT, particularly in technical fields. Instructor evaluations of student performance supported the findings, with fewer deficiencies in academic English skills for students with higher scores on TOEFL iBT. The study concludes that the TOEFL iBT's predictive validity is on par with the institution's own proficiency test and represents a solid performance measure for use in EMIs in a non-target-language context.

Keywords the TOEFL iBT® test; predictor; validity; academic performance

doi:10.1002/ets2.12230

Access to undergraduate university education in Turkey is decided on the basis of the results of a national university entrance exam designed and administered by the Student Selection and Placement Center. Students compete for a place in a university department of their choice based on their ranking on this national exam. Students' level of English is not assessed in the placement decision, although it figures in the national school curriculum from Grades 2 to 12. The five Turkish universities within the Times Education top 500[1] are English-medium institutions (EMIs). Many of the other 206 universities in Turkey also teach all or part of their programs in English.

Students who are placed in an EMI are required to demonstrate their proficiency in English upon registration. Proficiency is generally assessed by sitting for the EMI's in-house English proficiency exam or by submitting a valid score on a respected international exam, such as the *TOEFL iBT*® test or another accepted external test. Each university determines its own entry requirement on the TOEFL iBT. Thus crucial high-stakes decisions for access to freshman programs in EMIs are based on in-house tests or external measures, and establishing the predictive validity of English tests, based on a cut score at entry that gives students the greatest chance of success in their studies, represents a moral obligation on the part of such institutions, particularly in non-target-language communities (O'Dwyer & Atlı, 2018).

The study reported here was carried out in the School of English Language (SEL) at Bilkent University, Ankara, Turkey, which teaches most of its courses in English. The SEL's preparatory English program has approximately 2,500 full-time students in the main Turkish nationals (e.g., the majority of the students are Turkish). The program prepares students to meet the English-language requirements for access to their faculty courses and is assessed by an in-house proficiency exam covering the four skills plus use of language, called the Certificate of Proficiency in English (COPE), benchmarked to Level B2 on the Common European Framework of Reference (CEFR; Kantarcıoğlu, Thomas, O'Dwyer, & O'Sullivan, 2010). In addition, the SEL delivers credit-bearing English for academic purposes (EAP) freshman courses and other language courses in sophomore, junior, and senior years to some 3,000 students per semester. A Level B2 on the CEFR is considered a necessary but not sufficient condition for successful study in the medium of English; therefore, continued EAP support is required once students access the faculties. English preparatory and freshman EAP programs are situated

Corresponding author: E. Kantarcıoğlu, E-mail: kutevu@bilkent.edu.tr

within the SEL, which facilitates the integration of language curricula and offers the opportunity to do research over an extended period based on a knowledge of learner characteristics and language-learning history.

A predictive validity study of the TOEFL iBT provided an opportunity for the SEL to ensure that it was meeting its obligations toward students by achieving one of its key missions: to "ensure(s) that they attain the level of proficiency in English necessary to enter their chosen School or Faculty" (Bilkent University School of English, n.d.). The study also gave the possibility of further empirical validation of its own proficiency exam (COPE) relative to a respected international benchmark, the TOEFL iBT.

Literature Review

A number of predictive validity studies, as reported in Fox (2004), "find the investigation of the relationship between language tests and academic outcomes a futile line of inquiry" (p. 461), given, one presumes, the complexity of the task and the number of confounding variables. Others, however, consider test validation a crucial part of verifying a test's fitness for purpose, of which predictive validity is an essential element, given the high-stakes decisions for entry into English-medium academic courses taken on the basis of scores in key language tests (for complete validation models, see Mislevy, Steinberg, & Almond, 2003; Weir, 2005). The perspective adopted in this study supports the latter argument in recognizing the need to assess the impact of tests over time to "capture test-takers' abilities to function successfully in academic contexts" (Zareva, 2005, p. 47), despite the challenges this presents.

The literature on predictive validity studies is developing in terms of the number of studies undertaken and, perhaps more importantly, the methodologies they adopt. Academic performance has been operationalized in a good number of predictive validity studies through using a student's grade point average (GPA) as the outcome measure for academic success; this has then traditionally been correlated with test scores on a chosen English exam to yield a measure of the amount of the outcome variance attributable to a candidate's language proficiency level on the administered test. Fox, Cheng, Berman, Song, and Myles (2006) suggested extending this measure to include credits achieved as a more defensible criterion to predict; Lee and Greene (2007) pointed to the problems of comparison of studies due to differing measures of outcomes and proficiency. Notwithstanding, a number of studies have reported a weak positive correlation between academic success and language measures (Cho & Bridgeman, 2012; Feast, 2002; Fox, 2004; Hill, Storch, & Lynch, 1999; Lee & Greene, 2007; Manganello, 2011; Woodrow, 2006), with some disciplines having a more favorable predictive correlation, albeit still weak (Sawaki & Nissan, 2009; Wait & Gressel, 2009), and with some proficiency measures performing better than others.

Some studies have differentiated between lower scoring groups and higher scoring groups in an English test and then assessed academic performance (Cho & Bridgeman, 2012; Wait & Gressel, 2009). They found that a higher level of language ability on entry positively affects academic performance but that, over and above a certain level of language competence, it plays no significant role in academic performance. According to Elder (1993), it is at lower levels of proficiency that language makes a difference (corroborated by Fox, 2004; Woodrow, 2006), and it is there where students may require additional support (Hill et al., 1999). However, the evidence is inconclusive; "One can have a high TOEFL score and still experience a high level of academic difficulty" (Xu, 1991, p. 568), as language proficiency is but one of the requirements for academic success.

Several studies have looked at the predictive validity of subskills on proficiency tests and concluded that certain subskills of language proficiency might be better predictors of academic outcomes, suggesting minimum subskill scores in conjunction with an overall score (Golder, Reeder, & Fleming, 2009). Paul (2007) suggested that higher levels of language proficiency may help with academic task demands through reducing stress and fostering success; however, a higher language pass may not help if there is no attention to task and discourse demands of particular disciplines.

In few studies has the language development over and above the threshold level been measured, with the underlying assumption that this is part of the discipline-specific academic work, although a distinction is made between general academic language and discipline-specific academic language. In Fox (2004), those students who got higher than a score deemed sufficient for provisional acceptance, but not high enough for unconditional acceptance, continued their EAP course but were permitted to take a limited number of discipline courses; the system thus operated with two criterion-level scores, the higher level score eliminating the need for a student to get additional language support, allowing him or her to embark fully on his or her degree courses. Lloyd-Jones, Neame, and Medaney (2007) followed graduate students who had exhibited a borderline score, using several measures; over half the borderline students were instructed to revise and

resubmit their theses. The aggregate scores hid variations in subscores. The scarcity of attempts to measure the developing English-language skills of those who met threshold language requirements might be considered an oversight in predictive validity research. It seems reasonable to suggest that, once students are embarked on their academic courses, levels of English academic language competence should develop, if all courses are in the medium of English.

Quantitative studies have been undertaken in several different countries and university systems, often in contexts where students came to the target-language communities to study in universities in those countries, for example, in the United States (Cho & Bridgeman, 2012), Canada (Fox, 2004), the United Kingdom (Yen & Kuzma, 2009), and Australia (Woodrow, 2006). As a result, subjects in the studies were from a range of different nationalities arriving in a host country to continue their studies (Cope, 2010) or had been in the country up to 2 years or more in some cases prior to taking the language exam (Fox, 2004). Thus the subjects incorporated into studies were nonhomogenous in terms of background, although one U.K. study focused exclusively on Chinese students' academic performance (Yen & Kuzma, 2009) and one Australian study focused on Vietnamese learners (Huong, 2001). It appears that relatively few studies have taken place outside target-language communities, with a few exceptions; for example, a study in the United Arab Emirates, although it took place in a single location, still drew on students from many parts of the Arab world and, therefore, incorporated different nationalities into the sample (Wait & Gressel, 2009). A suggestion arose in conclusions from the analysis of the data that different nationalities (and genders) might have a different response to the academic demands of courses and, therefore, tantalizingly posited the need for more research into the relationship between nationality, language, and academic outcomes.

Several constraints may be noted in relation to quantitatively oriented studies. The duration of research reported can be seen to be short in some studies, focusing on academic achievement in the first one or two semesters in a freshman year; this means that academic success is not gauged by degree completion. Whether longer term considerations can be taken into account in relatively circumscribed studies is a moot point. Kokhan (2012) reported a year as being long enough for scores from the TOEFL iBT test to lose their relevance; in other words, TOEFL iBT scores were better predictors of short-term than long-term academic success, although the correlation recovers somewhat after a year. But limiting the period in which outcomes are assessed does lead to a potential hole in the longer term predictive impact of a language measure. This trend must be particularly worrying in studies in which only the first semester's GPA is used as a measure of academic performance or where GPA is gauged on assignment work (cf. Woodrow, 2006).

The heterogeneity of samples in many studies, due to having been undertaken in the target-language community, would seem to limit the possibility to control for factors such as school background or performance on university entry exams as predictors of academic success. The range restriction problem is also an issue in detecting predictive validity (Cho & Bridgeman, 2012). All students, particularly those failing to make the cut score, are not included in samples, which may dampen or underestimate correlation coefficients. According to Hirsh (2007), older studies of predictive validity may not be relevant, as data used were truncated because they looked at a population above a certain level of language proficiency only.

Many studies recognize that other factors are at work in academic performance (Dooey & Oliver, 2002; Elder, Bright, & Bennett, 2007), given the low variance attributed to language as a predictor of academic success. Some of these might be age (Xu, 1991); workload, concepts, resources, or teachers (Hill et al., 1999); sociocultural and psychological factors, learning and educational styles, motivation and maturity, or family and financial pressure (Kerstjens & Nery, 2000); adaptability to a new learning system, speed of acculturation, personal goals, ambition, or sociocultural factors (Yen & Kuzma, 2009); or motivation, learning strategies, or quantitative reasoning (Cho & Bridgeman, 2012). As Ingram and Bayliss (2007) concluded, "numerous variables intervene between proficiency and academic success" (p. 5), including intellectual ability, motivation, quality of teaching, learning style, and acculturation; therefore it is impossible to account for all variables. Some studies have suggested that students' own self-evaluation of their language levels is a better predictor of academic success (Dooey & Oliver, 2002; Xu, 1991), correlating closely with lecturers' perceptions (Cotton & Conrow, 1998).

Research Design and Methodology

Research Questions

The research literature documents limitations associated with predictive validity studies, suggesting the need for more comprehensive research designs, including studies with a homogenous population of learners within a

Table 1 Breakdown of the Components of the Institution's Proficiency Test, the Certificate of Proficiency in English

Section	Format and number of items	Time allocated	Weighting/150
Reading	35 multiple choice items	1 hour 20 min	35 points
Listening	Note taking then 30 multiple choice items after listening once to 2 lectures	1 hour	30 points
Writing	1 essay task with a choice of 2 prompts	1 hour	30 points
Speaking	Interview (one-on-one)	∼7 min	20 points
Language	20 open cloze items 7 gap-fill vocabulary items 8 word-formation items	40 min	35 points

non-target-language community, with performance measures differentiating discipline-specific and language course outcomes, with learners below a threshold proficiency level, and with a longitudinal approach covering a range of factors affecting academic success.

This study set out to research the predictive validity of the TOEFL iBT in a relatively homogenous student population with known demographics, namely, age, learning background, and university entrance scores, in a non-target-language community where acculturation was not a major issue. Evidence from multiple sources was collected: test scores on the TOEFL iBT, scores on the in-house proficiency test (COPE), scores from freshman English-language courses (ENG 101 and ENG 102), GPAs for all faculty courses for two freshman semesters, freshman student self-evaluations of their language performance, and EAP instructor evaluations of students over the freshman year. The study was able to include students who did not meet the TOEFL iBT requirements to enter the faculties, providing a less restricted range of available data. The latter students were admitted to faculties based on their results on the school's own proficiency test (COPE), despite not meeting the university's TOEFL iBT entry benchmark.

The research questions (RQs) for the study were as follows:

RQ1: What is the relationship between TOEFL iBT scores and future academic performance, as defined by GPAs?
RQ2: What is the relationship between TOEFL iBT scores and future academic performance, as defined by EAP course outcome measures in the freshman year?
RQ3: Does the relationship between the TOEFL iBT and future academic performance vary by discipline?
RQ4: How does the TOEFL iBT perform in relation to the institution's own English proficiency exam, COPE, both at an aggregate level and in relation to individual subskills?

Implementation of the Study

TOEFL iBT measures four skills — reading, listening, writing, and speaking — and follows an integrated approach to the testing of writing and speaking skills. The test is delivered via the Internet.[2] The institutional test, COPE, is a paper-based test set at CEFR B2 level (Kantarcıoğlu et al., 2010; Thomas & Kantarcıoğlu, 2009) and also measures the preceding four skills, with the addition of a language component. A breakdown of the components is given in Table 1.

A total of 658 students, in what is a mainly monolingual, non-target-language community, participated in the study. They were in the highest level of five consecutive levels (elementary, preintermediate, intermediate, upper intermediate, prefaculty) in the English-language preparatory program. International students were excluded. For scoring reliability, all components of the TOEFL iBT were graded externally by Educational Testing Service (ETS).

The software provided by ETS was set up on machines in university computer labs and tested prior to the project by the researchers and technicians. However, several problems were experienced during exam administration. First, pressing the "next" button closed the test program for some students, causing them to restart the test from the very beginning, leading to frustration and a number giving up. Roughly half of the first day's exam takers, approximately 175 students, experienced this. Second, the speaking section data files were not collected on the server for a number of students who had successfully completed the section, which became apparent only when students contacted the school to learn their scores. No data exist as to the number concerned. Finally, despite volunteering for the project, approximately 200 students decided to exit the test early.

Thus, of the 658 students who took the exam over 2 days, only 361 TOEFL iBT scores were available to the researchers, distributed as follows: 206 with scores on four sections; 121 with scores on three sections, with either a speaking score or a writing score missing; and 34 with scores on fewer than three sections. Only 56 students scored 80 or above on the

Table 2 Summary of Data Available for the Analysis

Scores	TOEFL	Passed COPE	Eliminated	Actual
TOEFL iBT scores on 4 sections	206	192	8	184
TOEFL iBTscores on 3 sections	121	121	16	105
TOEFL iBT <3 sections	34	34	34	–
Total	361	347	58	289

Note. COPE = Certificate of Proficiency in English.

TOEFL iBT, the score required by the university to pass directly into its faculty; only 42 achieved the (then) TOEFL iBT notional boundary for a B2 level, an average of 86 and above. Of the 361 TOEFL iBT scores available to the researchers, 347 students passed the institutional proficiency exam (COPE). Fifty-eight scores were eliminated from the data: 34 with scores on fewer than three TOEFL iBT sections and 24 others for bureaucratic reasons (e.g., transfers). A further three students took leaves of absence for the first year, leaving 286 data points. Within this number was a large cohort who scored below 86 on the TOEFL iBT, the (then) official B2 cut score (see Table 2).

Prior to the TOEFL iBT exam, 45 preparatory program instructors completed a questionnaire asking them to predict students' performance on COPE and TOEFL iBT by indicating either pass or fail. These same instructors gave TOEFL iBT familiarization lessons as part of the run-up to the study. Students were asked to complete a learner characteristics questionnaire at the start of familiarization and a short survey just after taking the TOEFL IBT. The freshman EAP course instructors answered two questionnaires: The first identified instructors' general perceptions of the academic English skills that students needed and if these skills were lacking; the second evaluated the performance of individual students using the same criteria. The students themselves were asked to evaluate their own performance, using the same criteria as their instructors. A questionnaire was sent to freshman content instructors; the response was poor, so content instructor perspectives do not figure in the analysis.

Correlations were computed for scores on TOEFL iBT, COPE, ENG 101, ENG 102, and first- and second-semester GPAs, both at an aggregate level and for the different sections of the TOEFL iBT and COPE tests. ENG 101 and ENG 102 letter grades were converted to a numerical scale (A – F converted to 1 – 10) for computing correlations. The two freshman-semester GPAs and English grades were added together to give composite totals for the year. Correlation coefficients were calculated separately for students with scores on four sections and those with scores on three sections of the TOEFL iBT. For some of the analysis, scores were weighted and combined. Correlations for scores broken down into academic disciplines were also computed. TOEFL iBT scores were compared with EAP instructor evaluations of students on ENG101 and ENG 102 courses and with students' self-evaluations.

Profile of TOEFL iBT Study Student Participants

The bio-data from the survey on learner characteristics, completed by 421 students prior to them taking the COPE and TOEFL iBT, revealed that 35% of students had started learning English around the age of 11 years and that 23% had started around the age of 8 years. The first language of all participants was Turkish. Prior to enrollment in the SEL's preparatory program, 40% had had between 3 and 5 hours of English per week, while 32% had had fewer than 3 hours per week; 46% had had experience in taking an international language exam. Almost all (94%) stated that it was their choice to study at an EMI where the level of English expected of students was high; 72% stated that they were motivated in their English classes.

Results

Student Feedback on Exam Administration Factors Impacting on the TOEFL iBT

Approximately 30 students took the TOEFL iBT simultaneously in each of five computer laboratories set up for the purpose. At the end of the test, 595 students completed a survey on the computer-based test experience; 31% ($n = 184$) of respondents indicated that using the Qwerty keyboard affected their performance (in Turkey, an F keyboard is common). Noise emanating mainly from students completing the speaking component (42%) and keyboards (31%) proved to be

Table 3 Correlations for Students With Scores on Four Sections of TOEFL iBT and Passing the Certificate of Proficiency in English

Scores	SEM 1 GPA[a]	SEM 2 GPA[a]	ENG 101[a]	ENG 102[b]
Aggregate TOEFL iBT scores	.36	.32	.37	.45
Aggregate COPE scores	.39	.38	.43	.43

Note. All correlations are significant at the level of $p < .01$; differences among TOEFL sections and COPE scores are not significant. GPA = grade point average; COPE = Certificate of Proficiency in English; SEM = semester.
[a]$N = 177$. [b]$N = 136$.

Table 4 TOEFL iBT and Certificate of Proficiency in English Scores Correlated With Freshman Grade Point Average and Combined English Scores

Section	TOEFL total	TOEFL LRW	TOEFL LRS	GPA total	ENG total
TOEFL total	–	–	–	.38[a]	.58[b]
TOEFL LRW	–	–	–	.37[c]	.50[d]
TOEFL LRS	–	–	–	.37[e]	.54[f]
COPE total	.81[g]	.78[h]	.80[i]	.46[a]	.57[b]

Note. All correlations are significant at the level of $p < .01$; differences among TOEFL iBT sections and COPE scores are not significant. GPA = grade point average; LRS = listening, reading, speaking; LRW = listening, reading, writing; COPE = Certificate of Proficiency in English.
[a]$N = 170$. [b]$N = 123$. [c]$N = 242$. [d]$N = 177$. [e]$N = 198$. [f]$N = 141$. [g]$N = 181$. [h]$N = 258$. [i]$N = 209$.

a distracting factor. Furthermore, 79% of the students opined that taking the exam on a computer affected their performance negatively. Only 24 students perceived a computer-based exam as positive, with 100 stating it did not affect their performance. The institutional exam, COPE, is paper based, so no such issues arose.

Teacher Estimates of Success on the TOEFL iBT and the Certificate of Proficiency in English

Teachers involved in the familiarization training prior to the TOEFL iBT estimated separately the likelihood of their students meeting minimum institutional TOEFL iBT requirements and of passing COPE. Of the 206 students with scores on four sections of the TOEFL iBT, teacher estimates were available for 198, and they were available for the 588 students who took the COPE. COPE estimates were largely accurate: From a predicted 381 passes, 321 passed (84%); from a predicted 207 failures, 166 failed (80%). TOEFL iBT predictions were less accurate: For available data, from a predicted 63 passes, 23 passed (37%); however, from a predicted 135 failures, 124 failed (92%). More difficulty was experienced in estimating around the (then) university cut score of 80 on the TOEFL iBT.

TOEFL iBT/Certificate of Proficiency in English Correlations With Grade Point Average and English 101 and English 102 Course Results

Table 3 presents correlations of the aggregate scores for those who completed four sections of the TOEFL iBT for those who completed four section and those who passed the COPE (cf. Table 2) with their Semester 1 and 2 GPAs and ENG 101 and ENG 102 grades. COPE scores exhibit marginally higher correlation coefficients on three of four performance measures, but none of these differences were statistically significant ($p > .05$).

Table 4 groups students who completed either four sections on the TOEFL iBT (TOEFL total) or listening, reading, and writing (TOEFL LRW) or listening, reading, and speaking (TOEFL LRS). The first- and second-semester GPAs were added together to give a combined GPA total score (GPA total). Similarly, ENG 101 and ENG 102 grades were summed to give a combined English total score (ENG total). The TOEFL total correlation with GPA total (.38) is marginally lower than the COPE total correlation (.46). Almost identical aggregate scores on both exams display moderately high correlations with success in freshman English courses. COPE total correlates highly with TOEFL total, TOEFL LRW, and TOEFL LRS.

Table 5 shows the correlation of individual sections on the TOEFL and COPE with the GPA total and the English total scores over the two freshman semesters. Reading and writing scores on both tests display moderate to moderately high correlation coefficients, as do TOEFL Listening scores.

Table 5 TOEFL iBT and Certificate of Proficiency in English Correlated With Grade Point Averages and English Total Scores Over Two Semesters

Scores	TOEFL Listening	TOEFL Reading	TOEFL Writing	TOEFL Speaking	COPE Language	COPE Listening	COPE Reading	COPE Writing	COPE Speaking
GPA total[a]	.31	.31	.29	.30	.46	.30	.31	.31	.29
ENG total[b]	.48	.53	.45	.37	.51	.35	.44	.46	.39

Note. All correlations are significant at the level of $p < .01$. GPA = grade point average; COPE = Certificate of Proficiency in English. [a]$N = 170$. [b]$N = 123$.

Table 6 Comparison of Top and Bottom TOEFL iBT and Grade Point Average Quartiles as a Percentage of Test Takers

	GPA quartiles	
TOEFL iBT quartiles	Bottom (<4.49)	Top (>6.80)
Top (>76)	2	23
Bottom (<52)	16	2

Note. GPA = grade point average.

Table 7 Correlations for Faculty on Aggregate TOEFL iBT and Certificate of Proficiency in English Scores

Faculty	SEM 1 CGPA	SEM 2 CGPA	ENG 101	ENG 102
Faculty of Art Design and Architecture				
TOEFL iBT total	.110 (33)	.232 (33)	.361* (33)	.259 (27)
COPE total	.190 (33)	.110 (33)	.480** (33)	.340 (27)
Faculty of Business Administration				
TOEFL iBT total	.294 (24)	.163 (24)	.185 (24)	.298 (18)
COPE total	.425* (24)	.318 (24)	.509* (24)	.187 (18)
Faculty of Economics Administrative and Social Sciences				
TOEFL iBT total	.199 (56)	.219 (56)	.068 (56)	.132 (41)
COPE total	.226 (56)	.255 (56)	.096 (56)	−.023 (41)
Faculty of Engineering				
TOEFL iBT total	.480** (97)	.459** (97)	.415** (97)	.366** (80)
COPE total	.594** (97)	.540** (97)	.450** (97)	.346** (80)
Faculty of Humanities and Letters				
TOEFL iBT total	.071 (9)	−.085 (9)	.294 (9)	.215 (8)
COPE total	.239 (9)	.263 (9)	.332 (9)	.709* (8)
Faculty of Law				
TOEFL iBT total	.196 (21)	.238 (21)	.008 (21)	.260 (17)
COPE total	.495* (21)	.532* (21)	.159 (21)	.392 (17)
Faculty of Science				
TOEFL iBT total	−.149 (15)	.096 (15)	−.297 (15)	.068 (12)
COPE total	.205 (15)	.441 (15)	.163 (15)	.190 (12)
School of Applied Technology and Management				
TOEFL iBT total	.011 (11)	−.107 (11)	−.419 (11)	−.621 (6)
COPE total	−.427 (11)	−.247 (11)	−.498 (11)	.140 (6)

Note. N is in parentheses. CGPA = cumulative grade point average; COPE = Certificate of Proficiency in English. SEM = semester. *$p < .05$. **$p < .01$.

Table 6 exhibits the performance of students who returned scores on four sections of TOEFL iBT in the top and bottom quartiles compared to the top and bottom quartiles of their GPA (GPA total) grades; GPA is out of four, here doubled, out of eight, to give a score for the two semesters. Top-quartile TOEFL iBT students were 10 times more likely to be in the top GPA quartile as in the bottom GPA quartile.

Table 7 reports correlations with GPAs and EAP scores calculated for students in different faculties based on aggregate scores on TOEFL iBT and COPE. Engineering faculty correlations are high relatively, with a number of meaningful results for law, business, and architecture.

Table 8 Correlations for Technical and Social Sciences Majors

	Technical fields		Social sciences	
Scores	TOEFL iBT total	COPE total	TOEFL iBT total	COPE total
GPA total	.53[a]	.60[b]	.35[c]	.42[d]
ENG total	.64[e]	.57[f]	.48[g]	.42[h]

Note. All correlations were significant at the level of $p < .01$. GPA = grade point average; COPE = Certificate of Proficiency in English. [a]$N = 67$. [b]$N = 115$. [c]$N = 103$. [d]$N = 155$. [e]$N = 51$. [f]$N = 87$. [g]$N = 72$. [h]$N = 108$.

Table 9 Freshman English Instructor General Perceptions of English for Academic Purposes Students and Evaluation of TOEFL iBT Test Takers

	Speaking		Writing		Reading		Listening	
Perceptions	SEM 1	SEM 2	SEM 1	SEM 2	SEM 1	SEM 2	SEM 1	SEM 2
Needed	.86	.86	.87	.95	.71	.93	.85	.97
Deficiencies	.63	.51	.66	.47	.61	.54	.47	.41
TOEFL iBT average deficiencies	.37	.32	.40	.40	.28	.33	.32	.34

Note. SEM = semester.

Students with scores on all four TOEFL iBT sections were allocated to one of two general categories: science, computer technology, engineering, and mathematics majors or social sciences–related fields. Results in Table 8 suggest that the TOEFL iBT and the COPE are stronger predictors of performance in technical fields for both content and EAP courses.

English for Academic Purpose Freshman Questionnaire Data

Questionnaires were given to ENG 101 and ENG 102 EAP instructors to collect data on TOEFL iBT students' English-language performance. Students were assessed on 25 skills-related statements (eight on speaking, three on writing, eight on reading, six on listening) to get instructor perceptions of students' needs and deficiencies, taken as a whole, in their ENG 101 and ENG 102 sections. Thirteen ENG 101 instructors and 15 ENG 102 instructors returned questionnaires. Perceptions of student needs and deficiencies were scored numerically, with yes = 1 and no = 0 for each of the 25 statements, then averaged for each skill category for each instructor, and then averaged for all responding instructors.

Instructors also evaluated the needs and deficiencies of individual students who had taken the TOEFL iBT, using the same 25 statements. Forty-five forms for ENG 101 students were returned in Semester 1, and 126 ENG 102 forms were returned in Semester 2. Some of the Semester 1 students were present among the 126 students in Semester 2. For the purposes of the analysis in Table 9, the scores for students with results in three TOEFL iBT sections were divided by 3 and multiplied by 4 to give a score out of 120. The resulting scores were grouped with those students with scores on four sections. Table 9 presents the averaged scores: Closer to 0 indicates less needed or less deficient, and closer to 1 indicates more needed or more deficient. These scores may be read as percentages, for example, .86 and .63 for Semester 1 (speaking) in Table 9 may be interpreted as 86% of instructors considered the skill needed by students, and 63% of the students were deficient in this skill.

The data show an increase in EAP instructor expectations in writing, reading, and listening skills in Semester 2, consistent with a change in emphasis in the EAP courses between ENG 101 and ENG 102. ENG 102 students are required to produce a research-based term paper that is presented orally to a committee of EAP instructors as part of the final course assessment. Deficiencies appear lower in Semester 2, which suggests that instructors considered all students to have made progress in the skills, more noticeably in the productive skills, particularly writing. Individual students in EAP classes who had higher TOEFL iBT scores had fewer deficiencies on average in the view of EAP instructors.

Table 10 compares the deficiencies of students who scored higher than 70 on the TOEFL iBT with those who scored 70 or below. The data show deficiencies, based on instructor perceptions, between the higher scoring and lower scoring students.

Table 10 Instructor Questionnaire Analysis for Higher and Lower Scoring Students on the TOEFL iBT

Perceptions	Speaking	Writing	Reading	Listening
Student deficiencies, TOEFL score ≥70[a]	.21	.27	.20	.26
Student deficiencies, TOEFL score < 70[b]	.35	.45	.34	.38

[a]$N = 62$. [b]$N = 122$.

Table 11 Student Confidence Levels in Skills in Freshman Courses

	Speaking		Writing		Reading		Listening	
Level	SEM 1	SEM 2	SEM 1	SEM 2	SEM 1	SEM 2	SEM 1	SEM 2
Confidence level[a]	3.12	3.43	3.29	3.41	3.57	3.60	3.29	3.56
N	44	121	44	121	44	121	44	121

Note. SEM = semester.

[a]1 = very low. 2 = low. 3 = medium. 4 = high. 5 = very high.

Table 12 Semester 1 Instructor Perceptions of Individual Student English for Academic Purposes Skills for TOEFL iBT ≥70

n	TOEFL iBT score	TOEFL iBT average score	Speaking	Writing	Reading	Listening
13	≥90	94.15	0.23	0.49	0.24	0.14
13	80–89	84.15	0.33	0.24	0.20	0.29
16	70–79	74.08	0.55	0.43	0.38	0.57

A student self-evaluation questionnaire asked students to rate themselves on the same skill areas as were present in the instructor questionnaire. Forty-four ENG 101 and 6 ENG 102 students responded. The student confidence levels illustrated in Table 11 indicate that aggregate confidence levels in the four skills improved to a degree in the second freshman semester.

Instructor perceptions of deficiencies were calculated for different bands of TOEFL iBT scores, using averaged scores on three and four TOEFL iBT sections. Tables 12 and 13 break down the deficiencies relative to the levels achieved on the TOEFL iBT. Table 12 shows that Semester 1 students who scored higher than 80 tended to have fewer deficiencies in the eyes of the instructors. Those in the lower band generally had more deficiencies.

For Semester 2, a bigger range of bands was available. A decision taken in the first semester to restrict analysis to only those who scored 70 and higher on the TOEFL iBT was changed in the second semester to include all TOEFL scores available. Table 13 includes both three- and four-section-scoring students and shows that those who scored lower on the TOEFL iBT had greater deficiencies in the eyes of the EAP instructors. The dividing line, at a glance, seems to be around 70 and higher for the TOEFL iBT. This finding suggests that those scoring higher than 70 on the TOEFL iBT were clearly better able to manage their second-semester freshman English courses than those who scored below this score.

Discussion

The study was undertaken in an English-medium university with a relatively homogeneous population in a non-target-language community, characteristics rare in the literature (Cho & Bridgeman, 2012). After taking the TOEFL iBT, students were tracked over their two freshman semesters. The institutional proficiency exam, COPE, was taken 10 days after the TOEFL iBT. Most students in the study would not have been admitted to faculties if their TOEFL iBT scores had been the only arbiter. Success rates in the COPE allowed the researchers to address the issue of range restriction, missing from most predictive validity studies (Hirsh, 2007).

A short training course was given to students prior to the TOEFL iBT to familiarize them with the format of the online exam. However, out of the 588 students who took the institutional test (COPE), only 209 completed four sections on the TOEFL iBT, and 121 completed three sections (Table 1); 34 returned two or fewer sections on the exam, and for the remaining 227, no scores were available. Reasons for this vary, among them a lack of familiarity with computer-based

Table 13 Semester 2 Instructor Perceptions of Individual Student English for Academic Purposes Skills for All TOEFL Scores

n	TOEFL	TOEFL average score	Speaking	Writing	Reading	Listening
9	≥90	95	0.03	0.00	0.04	0.06
11	80–89	86	0.18	0.39	0.16	0.29
11	70–79	75	0.28	0.18	0.22	0.23
35	60–69	65	0.37	0.50	0.36	0.32
26	50–59	56	0.30	0.39	0.31	0.37
31	≤49	39	0.45	0.56	0.51	0.52

testing: 79% of the 595 respondents to the post–TOEFL iBT survey reported that taking a computer-based test affected their performance to some degree. The Qwerty keyboard was also perceived as an issue, as was noise near the test takers, as 30 students took the exam simultaneously in a computer lab. The proximity to the institution's test (COPE), which students were more familiar with and which bestowed the same rights of passage on the test takers, may have led them to give up more easily on the TOEFL iBT. Technical difficulties arose with recording exam performance on the server or with software. Motivational issues among students may also have influenced the completion rate, as almost 30% of the exam takers reported themselves to be unmotivated prior to taking the exam. Thus, at the outset, exogenous factors, unrelated to language performance, impacted the predictive power of the study.

The data analysis provides clear evidence of the predictive validity of the TOEFL iBT for the academic performance of a homogenous population of learners in English-medium contexts outside the target-language community, the focus of the first RQ. The correlation of aggregate scores on TOEFL iBT with GPAs for two semesters was moderate, but respectable, at .38 (Table 4). Language skills represent only one part of the skills needed to succeed in academic course work, given the plethora of potential intervening variables (see "Literature Review"). In addressing the third research question, moderately high correlations for students studying technical discipline majors ($r = .53$) point to TOEFL iBT as an effective predictor of academic success for students in these fields (Table 8), although even for nontechnical majors, correlations remain moderately positive ($r = .35$). Certain disciplines might have more challenging discourse demands (Paul, 2007), which may explain the lower correlations in the social sciences, clearly an area ripe for further research.

Wait and Gressel (2009) found that TOEFL iBT test scores above a certain level play no significant role in predicting academic success. Instructor evaluations of individual student deficiencies (Table 10) show that higher scorers on TOEFL iBT have, on average, fewer English academic skill deficiencies. This would translate, ceteris paribus, into higher academic success in content courses as the analysis of TOEFL iBT and GPA quartiles bears out. Top-quartile TOEFL iBT students were 10 times more likely to be in the top GPA quartile (Table 6). Table 13 provides evidence that those averaging 95 on the TOEFL iBT are seen in a different light after two semesters by EAP instructors when compared to those averaging 85 and 75. In the same data set, students with a score of 70 and above, after two semesters of EAP, improved their language skills when compared to first-semester performance (Tables 12 and 13), whereas students with scores of less than 70 were evaluated by EAP instructors as skill deficient after two semesters, particularly in writing skills. As students scoring higher than 90 showed comparatively superior performance after two semesters, it can be argued that their higher TOEFL iBT scores played a significant role in predicting academic success, with a delay. Students who averaged 75 on the TOEFL iBT showed improvement in their EAP course performance in their second-semester results. EAP instructor evaluations for those students with scores around 75 were virtually indistinguishable from evaluations for those who averaged scores of 85. Lower scorers showed themselves capable of compensating for language deficiencies over time. At the time of data collection, the official TOEFL iBT B2 cut score was an aggregate of 86 on four sections, adjusted downward to 72 (Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015). The data suggest that the adjustment was well founded. In summary, very high-scoring TOEFL iBT students perform better on performance measures after a time delay, which suggests that very high scores do predict higher level academic outcomes for higher scoring than lower scoring students. However, over time, students with lower initial scores may be able to perform on par with those in a higher scoring band, suggesting that the predictive validity of scores is not written in stone above a well-delineated cut store.

The TOEFL iBT predicts, in response to the second research question, the academic performance of students on their EAP courses (ENG 101 and ENG 102). Test results correlate moderately well with performance on the two EAP courses taken separately (Table 3): first-semester $r = .37$; second-semester $r = .43$. When a combined EAP score over two semesters was calculated (Table 4), the correlation was moderately high ($r = .58$). These correlations provide further support for

improvement over time in the test's ability to predict success in English-language courses. However, the breakdown of faculty scores, despite the lack of significant correlations across the board, provides a caveat to this (Table 7). A reduction in second-semester EAP course correlations with their TOEFL iBT is apparent for Faculty of Engineering students. Unfortunately, no other statistically significant data exist for ENG 102 correlated with TOEFL iBT scores, but the institutional exam, COPE, shows a high significant correlation with Faculty of Humanities and Letters students. This finding may suggest that academic course demands in content courses in social sciences have a positive effect on ENG 102 outcomes, and the inverse for technical fields. Given the lack of significant data, the jury remains out on this question, another area ripe for further research.

The fourth research question targeted the concurrent and predictive validities of the institutional test when compared to the TOEFL iBT. Correlations of aggregate scores on both exams with GPA and EAP resemble one another (Table 3); the differences among TOEFL sections and COPE scores were not significant, with high correlations ($r = .8$) for aggregate scores on the TOEFL iBT test and COPE, whether returning three- or four-section scores on the TOEFL iBT (Table 4). The different sections on the TOEFL iBT and COPE (Table 5) correlate almost identically with combined GPA. Interestingly, the COPE language section has a similar correlation with the overall correlation of COPE with combined GPA (GPA in Table 4). TOEFL iBT and COPE section correlations with performance on combined English scores show that TOEFL iBT Reading and COPE Language carry more weight, although, again, differences are nonsignificant. A similar picture is apparent for technical fields versus social sciences in Table 8. COPE also shows itself to have predictive validity for the cohort of students who were the object of the study.

Conclusion

While acknowledging the limitations of the study, the authors believe that it contributes positively to the existing literature on the predictive validity research on the TOEFL iBT. It addresses the issue of range restriction and utilizes EAP instructor insights on student performance post-TOEFL iBT. Previous research relied in many cases on GPA as a performance measure. As the preceding discussion testifies, TOEFL iBT predicts moderate performance in content courses in the freshman year, with higher correlations achieved for performance in EAP courses. The results suggest that other factors enter into play when predicting performance on the academic content courses. The data obtained did not allow for an in-depth analysis of TOEFL iBT's predictive power for different faculties. The ideas put forward that academic content course demands in the social sciences might explain weaker GPA scores and, conversely, that social science course demands may heighten performance in EAP courses are interesting ones as a focus for further research. Unfortunately, the study fell short in terms of obtaining questionnaire data from departmental content course instructors. In addition, few one-on-one interviews were carried out with students to investigate their experiences with the TOEFL iBT and further explore how language skills measured were relevant to their academic studies.

The small size of the usable sample of TOEFL scores, despite the initial number who took the test, was a disappointment and prefaces the need for replication studies. However, in considering further research of this nature, the study points to the logistical issues raised when implementing a large-scale online examination within a restricted time of 2 days. EMIs in the Turkish context generally require a quick turnaround for tests of language given the short time between being placed and registered at a university and the start of courses. Large cohorts of students tested in computer labs may find validity compromised if results cannot be processed efficiently and effectively for one reason or another. Alternative means should be considered to administer TOEFL iBT to large numbers in a restricted time frame; otherwise, TOEFL iBT, for all its potential as a predictive measure, may not be a viable option for the university context.

Notes

1 See https://www.timeshighereducation.com/world-university-rankings
2 See https://www.ets.org/toefl/ibt/about/

References

Bilkent University School of English. (n.d.). *Mission statement*. Retrieved from http://busel.bilkent.edu.tr/?page_id=375

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29*, 421 – 442. https://doi.org/10.1177/0265532211430368

Cope, N. (2010). *Evaluating locally-developed language testing: A predictive study of direct entry language programs at an Australian university*. Macquarie Park, Australia: Centre for Macquarie English, Macquarie University.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports, 1*, 72 – 115.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS test as an indicator of future academic success. *Prospect-Adelaıde, 17*(1), 36 – 54.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing, 2*(1), 72 – 95.

Elder, C., Bright, C., & Bennett, S. (2007). The role of language proficiency in academic success: Perspectives from a New Zealand university. *Melbourne Papers in Language Testing, 12*(1), 24 – 58.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal, 3*(4), 70 – 85.

Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing, 21*, 437 – 465. https://doi.org/10.1191/0265532204lt292oa

Fox, J., Cheng, L., Berman, R., Song, X., & Myles, J. (2006). Costs and benefits: English for academic purposes instruction in Canadian universities. *Carleton Papers in Applied Language Studies, 23*, 1 – 108.

Golder, K., Reeder K., & Fleming, S. (2009). Determination of appropriate IELTS band score for admission into a program at a Canadian post-secondary polytechnic institution. *IELTS Research Reports, 10*, 69 – 94.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports, 2*, 52 – 63.

Hirsh, D. (2007). English language, academic support and academic outcomes: A discussion paper. *University of Sydney Papers in TESOL, 2*, 193 – 211.

Huong, T. T. (2001). The predictive validity of the international English language testing system (IELTS). *Post Script*, *2*(1), 66 – 96.

Ingram, D. E., & Bayliss, A. (2007). IELTS as a predictor of academic language performance. Part 1: The view from participants. *IELTS Impact Studies, 7*(3), 1 – 68.

Kantarcıoğlu, E., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). The COPE linking project: A case study. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual* (pp. 102 – 116). Cambridge, England: Cambridge University Press.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports, 3*, 85 – 108.

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*, 291 – 308. https://doi.org/10.1177/0265532211429403

Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*, 366 – 389. https://doi.org/10.1177/1558689807306148

Lloyd-Jones, G., Neame, C., & Medaney, S. (2007). A multiple case study of the relationship between the indicators of students' English language competence on entry and students' academic progress at an international postgraduate university. *IELTS Research Reports*, *11*, 129 – 184.

Manganello, M. (2011). *Correlations in the new TOEFL era: An investigation of the statistical relationships between IBT scores, placement test performance, and academic success of international students at Iowa State University* (Unpublished doctoral dissertation). Iowa State University, Ames, Iowa.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3 – 67. https://doi.org/10.1207/S15366359MEA0101_02

O'Dwyer, J., & Atlı, H. H. (2018). ESP/EAP in university programs in a non-target language community: Issues and challenges. In Y. Kırkgöz & K. Dikilitaş (Eds.), *English language education*: Vol. *11*. Key issues in English for specific purposes in higher education (pp. 291 – 304). Cham, Switzerland: Springer.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Paul, A. (2007). IELTS as a predictor of academic language performance, part 2. *IELTS Research Reports, 6*, 205 – 240.

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (TOEFL iBT Research Report No. TOEFLiBT-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02159.x

Thomas, C., & Kantarcıoğlu, E. (2009). Bilkent University School of English Language COPE CEFR linking project. In N. Figueras & J. Noijons (eds.), *Linking to the CEFR levels: Research perspectives* (pp. 119 – 124). Retrieved from www.ealta.eu.org/documents/resources/Research_Colloquium_report.pdf

Wait, I., & Gressel, J. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education, 98*, 389 – 398. https://doi.org/10.1002/j.2168-9830.2009.tb01035.x

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney Papers in TESOL, 1*, 51–70.

Xu, M. (1991). The impact of English-language proficiency on international graduate students' perceived academic difficulty. *Research in Higher Education, 32*, 557–570. https://doi.org/10.1007/BF00992628

Yen, D., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching, 3*, 1–7.

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

Suggested citation:

O'Dwyer, J., Kantarcıoğlu, E., & Thomas, C. (2018). *An investigation of the predictive validity of the TOEFL iBT® test at an English-medium university in Turkey* (TOEFL Research Report No. RR-83). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12230

Action Editor: John Norris

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFL Committee of Examiner