ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-95
ETS Research Report No. RR–21-19

New Validity Evidence on the *TOEFL Junior*® Standard Test as a Measure of Progress

Irshat Madyarov
Vahe Movsisyan
Habet Madoyan
Irena Galikyan
Rubina Gasparyan

December 2021

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

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

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

RESEARCH REPORT

New Validity Evidence on the *TOEFL Junior*® Standard Test as a Measure of Progress

Irshat Madyarov[1], Vahe Movsisyan[2], Habet Madoyan[3], Irena Galikyan[3], & Rubina Gasparyan[4]

1 Manoogian Simone College of Business and Economics, College of Humanities and Social Sciences, Yerevan, Armenia
2 Zaven and Sonia Akian College of Science and Engineering, Yerevan, Armenia
3 College of Humanities and Social Sciences, Yerevan, Armenia
4 American University of Armenia, Yerevan, Armenia

The *TOEFL Junior*® Standard test is a tool for measuring the English language skills of students ages 11+ who learn English as an additional language. It is a paper-based multiple-choice test and measures proficiency in three sections: listening, form and meaning, and reading. To date, empirical evidence provides some support for the construct validity of the TOEFL Junior Standard test as a measure of progress. Although this evidence is based on test scores from multiple countries with diverse instructional environments, it does not account for students' instructional experiences. The present paper aims to provide additional evidence by examining the TOEFL Junior Standard test as a progress measure within the same instructional setting. The study took place in an after-school English program in Armenia, a non-English-speaking country. A total of 154 adolescents took the TOEFL Junior Standard test three times with different test forms at the intervals of 10 and then 20 instructional weeks (a total of 30 weeks). The difference in differences (DID) analysis shows that TOEFL Junior is sensitive to learning gains within 20 instructional hours per 10 weeks among A1–A2 level learners, according to the Common European Frame of Reference (CEFR) scale. However, the data did not provide support for this sensitivity among B1–B2 level learners even though their instructional time was twice as long. Although this methodology offers an improved control over the students' instructional experiences, it also delimits the results to a specific after-school program and comes with a set of other limitations.

Keywords TOEFL Junior Standard; English as a foreign language; progress; intensity; Armenia; after school program; validity

doi:10.1002/ets2.12334

The TOEFL Junior Standard test is a tool for measuring the English language skills of students ages 11+ who learn English as an additional language (ETS, 2020). It is a paper-based multiple-choice test and measures proficiency in three sections: listening (42 questions, 40 min long), form and meaning (42 questions, 25 min long), and reading (42 questions, 50 min long). The test measures adolescents' ability to use English for interpersonal and academic purposes in English-medium instructional contexts, regardless of whether the contexts (ETS, 2020; So et al., 2015) are English as a second language (ESL) or English as a foreign language (EFL). The total resulting score can range from 600 to 900, with each section ranging from 200 to 300 points. The score report also shows equivalent levels of the Common European Framework of Reference (CEFR) for each test section, ranging from below A2 to B2 as well as *can-do statements* describing English language abilities typical of the given CEFR level (So et al., 2015).

Although the TOEFL Junior Standard test is the result of a rigorous process based on current empirical evidence, test validation is typically an ongoing process of accumulating evidence to facilitate proper interpretation and use of test scores. This kind of evidence concerns the construct validity of the test. So et al. (2015) provided a detailed discussion of the constructs that underlie the TOEFL Junior Standard test and their operationalization. An important framework that informs this test is Bachman and Palmer's (2010) model of language knowledge. This model assumes that language growth is due to different learning experiences, among other factors (e.g., Chapelle et al., 2008; Messick, 1989).

There is some empirical evidence suggesting a relationship between English learning experiences and change in students' performance measured by standardized tests. For the International Language Testing System (IELTS), these results suggest that academic English gains are achievable within 3–12 weeks of instruction, on average ranging from 0.21 to half a band, measured on a scale of 0–9 and reported in increments of 0.5 points (Elder & O'Loughlin, 2003; Green, 2004; Green & Weir, 2003). Learners at lower proficiency levels (IELTS global scores of 3.0 through 5.0) tend to show larger gains

Corresponding author: I. Madyarov, E-mail: imadyaro@aua.com

than those scoring 6.0 and higher (Elder & O'Loughlin, 2003; Green, 2004; Green & Weir, 2003). For example, students with 3.0 on IELTS at pretest were able to gain more than 1.5 points in writing within 3 months of instruction (Green & Weir, 2003). However, students with 7.0 at pretest tend to score lower at posttest even after a 12-month period of instruction. In other words, this evidence suggests that the higher the proficiency level of the learner, the harder it becomes to make larger academic English gains.

Returning to the TOEFL family of assessments, Ling et al. (2014) demonstrated that the *TOEFL iBT*® is capable of capturing language progress over 6–9 months of intensive study of English in the United States and China among high school students ($N = 111$). They used a pre- and posttest design for a subset of their data based on the TOEFL iBT practice test. Wolf and Steinberg (2011) used a cross-sectional design to compare the performance of US middle school students, both native speakers and language learners ($N = 2{,}622$), on the TOEFL Junior Standard test. Their results also suggest that the test can be used to monitor student progress in English-medium instructional contexts over time.

The study of Gu et al. (2015) occupies a special place in this review of literature. They conducted a longitudinal repeated measures design across 15 countries and regions among 4,606 TOEFL Junior Standard test takers. The number of times the test was taken varied from two to seven, and the duration between retesting ranged from 75 to 250 days. Although their data were rich in terms of country representation, instructional contexts, and duration between retests, they were not able to document students' specific instructional experiences, that is, whether the test takers had attended any English learning programs between the retests, and if so, what those programs were like. As such, the researchers relied on "the time interval between test administrations […] as a proxy for the amount of English language learning opportunities" (p. 3). Their results suggested that the TOEFL Junior Standard test is able to show changes in language performance due to the time factor, i.e., the longer the interval between retesting, the greater gains the test takers demonstrate. In their review of literature, Gu et al. (2015) also suggested that validation research with longitudinal designs to evaluate the appropriateness of score interpretation and use of the TOEFL Junior Standard test is scarce. In fact, our review of literature suggests that Gu et al. is the only such study with a retest design, and as the researchers acknowledged, it did not control for the students' instructional experiences.

The present study aims to address this gap by examining the TOEFL Junior Standard test as a progress measure relative to the English language learning experiences of students from the same after-school program. To this end, the study addresses the following research question: To what extent can the TOEFL Junior Standard test show gain, both in terms of overall proficiency and individual language skills (i.e., listening, reading, form and meaning) with learners of A1–A2 and B1–B2 levels studying English with different degrees of intensity (2 hr per week versus 4 hr per week)?

Method

Context

The after-school English program where the study took place offers extracurricular English classes to children ages 6 through 16 residing in the capital city Yerevan, Armenia, a post-Soviet country in the Caucasus region. This after-school program is part of the American University of Armenia and serves as a model school where teaching English as a foreign language (TEFL) student teachers carry out their teaching practicums and research. At the time of the study, the after-school program enrolled around 500 children in multiple classes of 10–16 children in each, grouped according to a placement test. Learners at the same level used the same ESL/EFL textbooks (published by Pearson and Cambridge), but the teachers had the freedom to supplement the textbooks with additional materials of their own choice. The teachers were Armenian-native speakers with master's degrees in TEFL from the American University of Armenia, an English-medium university.

The after-school program runs throughout the year except for 1.5 months in July and August. The curriculum spans 10 years, with 75–80% of student retention from year to year. New students can join the program at any school age and proficiency level. Beginning and elementary learners (pre-A1 through A2, CEFR) attend 1-hr classes twice a week for a term of 10 weeks. B1–B2 level learners and above attend 2-hr classes twice a week during the same 10 weeks. Thus, within a term, A1–A2-level learners complete a total of 20 hr of face-to-face instruction, and B1–B2 level learners complete a total of 40 hr of instruction. Most classroom instruction happens in English with minimal support in L1 for complex vocabulary. Communicative language skills are emphasized, with occasional explicit explanation of grammar for adolescent learners. Students are routinely assigned homework from workbooks or project-based tasks.

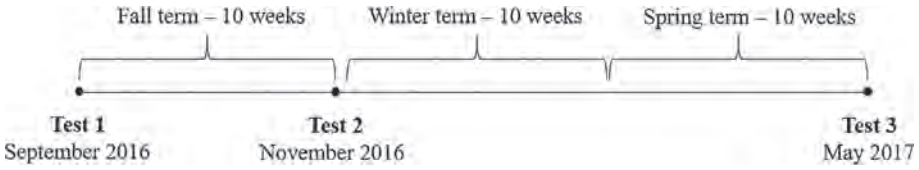

Figure 1 Timeline of test administrations throughout the total period of 30 instructional weeks.

Apart from attending the after-school program, almost all students also attended English classes in public or private secondary schools. Their weekly hours of instruction typically varied from 2 (in public schools) to 4 (in some private schools). It is unlikely that students in this study also participated in private tutoring classes outside school, because parents typically place their children in either after-school or private tutoring classes, but rarely in both. It is also unlikely that any of the student participants in this study were familiar with or preparing for the TOEFL Junior Standard test because this was and still is a new and unknown test in Armenia. No testing center in the country offers this test to the population, to our knowledge. All students in the target age group and their parents were invited to participate in the study. Those who agreed were asked to sign the TOEFL Junior Standard consent form.

Data Collection and Analysis

To collect data, the TOEFL Junior Standard test was administered three times, with a different form for each test administration (i.e., a single different form was used each time). The first administration of the test took place in late September (before the fall term). The second and third test administrations took place in late November (before the winter term) and early May (after the spring term), respectively. As a result, the intervals between testing times varied from 10 weeks (in the fall) to 20 weeks (winter + spring terms), with a total of 30 weeks between the start of the academic year in September and its end in May (see Figure 1). There were breaks after each term including the longest one of 3 weeks over the holidays in December and January.

The following are the sample sizes of the total number of students in the program, ages 11 – 17, in each term of the study: fall $n = 2$ 37 out of 345 students (69%), winter $n = 322$ out of 429 students (75%), and spring $n = 272$ out of 350 students (78%). Because the analysis included only those students who took the test three times, the resulting sample consisted of 154 students. The demographic characteristics of this sample are representative of the after-school student population in terms of age, gender, and proficiency level. Table 1 summarizes the student participants in the study by age.

Figure 2 shows that the distributions of the first test scores of the two programs (20 and 40 hr) were understandably quite different because the students were placed into classes according to a placement test. There is still some overlap of initial scores due to a few reasons. First, the placement test is specific to the textbooks used in the curriculum, and the TOEFL Junior scores cannot perfectly match curriculum-specific placement test results. Second, after the students are placed into their classes according to the placement test, the teachers can propose changes to the student placements based on their observations of the students' performance in class. This overlap of student scores in the 20- and 40-hr programs has implications for the study and will be discussed in the results and discussion sections.

All students whose parents signed the informed consents were invited to a large auditorium on the same day at the same time and seated with two vacant seats in between each student to prevent cheating. The ratio of students and proctors was

Table 1 Distribution of Students Who Took the Test Three Times, by Age

Age	Frequency	Percent
11	14	9.3
12	26	17.1
13	32	21.0
14	41	26.8
15	34	22.0
16	4	2.9
17	2	1.0
Total	154	100.0

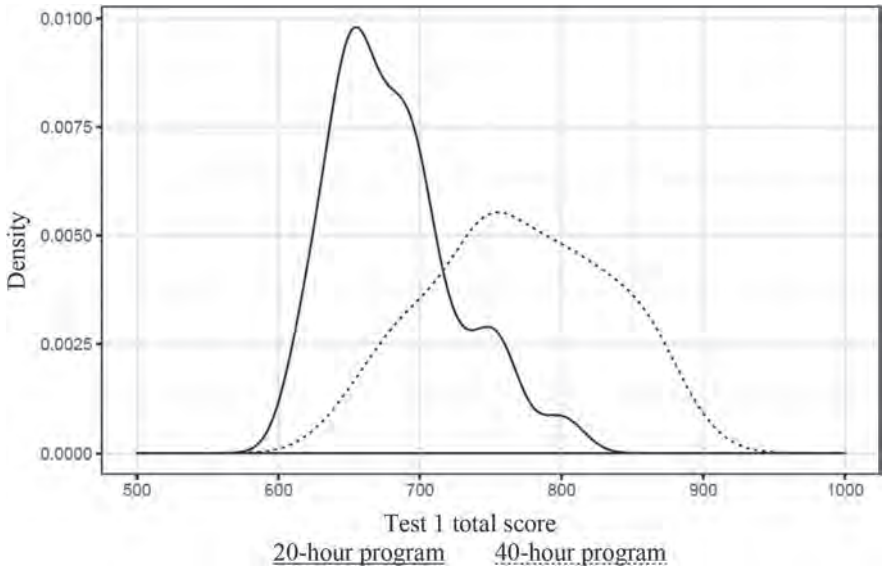

Figure 2 Distributions of the first test scores by program: <A2 – A2: 20 hr per term and B1 – B2: 40 hr per term.

approximately 20:1, in addition to a supervisor who led the testing sessions. The testing procedure followed the prescribed steps in the ETS manual for the TOEFL Junior Standard test.

The test data were analyzed in R, a programming language, which is also statistical software for computing and graphics. Descriptive statistics included frequencies, percent, means, and standard deviations for summary tables and graphs. To address the main research question, the difference in differences (DID) analysis of two factors was used: (a) the difference in time (Test 1, Test 2, Test 3) and (b) the difference in the program (20 hr, 40 hr). The model specification allows us to control for the initial difference of scores in the two groups and investigate the effect of the two programs. One limitation of this analysis in the study is that the duration between the tests and the length of the programs are not measured in continuous but discrete scales. This means that it was not possible to estimate elasticity of students' gains in increments of 1 hr of instructional time. Another limitation is that the initial test scores (Test 1) are not homogeneous across students, which makes the comparison across time more challenging. We discuss these challenges and solutions in the sections that follow.

Results

This study addresses the following research question: To what extent can the TOEFL Junior Standard test show gain, both in terms of overall proficiency and individual language skills (i.e., listening, reading, form, and meaning) with learners of A1 – A2 and B1 – B2 levels studying English with different degrees of intensity (2 hr per week versus 4 hr per week)?

Descriptive Statistics

Figures 3 through 6 summarize the students' progress measured at different intervals. One clear pattern from this visual summary is that there is an overall upward progression of scores between Time 1 and 3 for both A1 – A2-level students in the 20-hr program and B1 – B2-level students in the 40-hr program. On closer inspection, the listening scores do not follow this pattern between testing Time 2 (Nov-16) and Time 3 (May-17). This pattern is observable in both programs. This irregularity may seem particularly puzzling considering that the interval between testing Time 2 and Time 3 was 20 instructional weeks, twice as long as that between testing Time 1 (Sept-16) and Time 2 (Nov-16). In the discussion section, we offer a few possible explanations for this irregularity.

Tables 2 and 3 provide additional descriptive statistics. Table 3 shows a more detailed picture of the difference in the listening scores between Time 2 and Time 3 (0 points for the 20-hr program and −1.09 points for the 40-hr program). Overall, there is no clear pattern as to whether the 20-hr or 40-hr program demonstrates greater gains over the same periods of time. The DID analyses that follow provide some insights on the differences between the two programs.

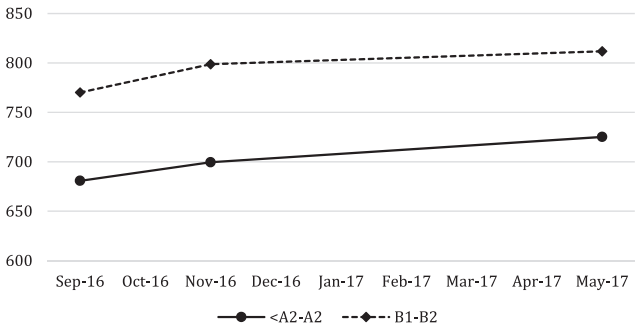

Figure 3 Students' progress in total scores measured at three testing times marked as total 1, 2, and 3. The progress data are presented by proficiency level and instructional time (A1 – A2: 20 hr per term; B1 – B2: 40 hr per term).

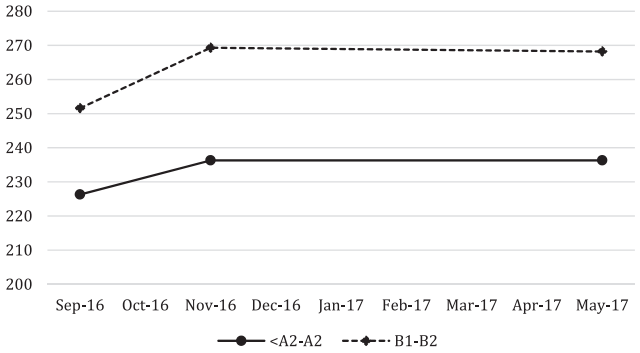

Figure 4 Students' progress in listening scores measured at three testing times marked as listening 1, 2, and 3. The progress data are presented by proficiency level and instructional time (A1 – A2: 20 hr per term; B1 – B2: 40 hr per term).

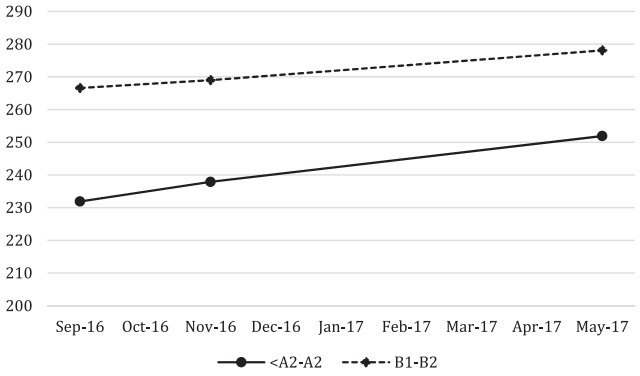

Figure 5 Students' progress in form and meaning scores measured at three testing times marked as form and meaning 1, 2, and 3. The progress data are presented by proficiency level and instructional time (A1 – A2: 20 hr per term; B1 – B2: 40 hr per term).

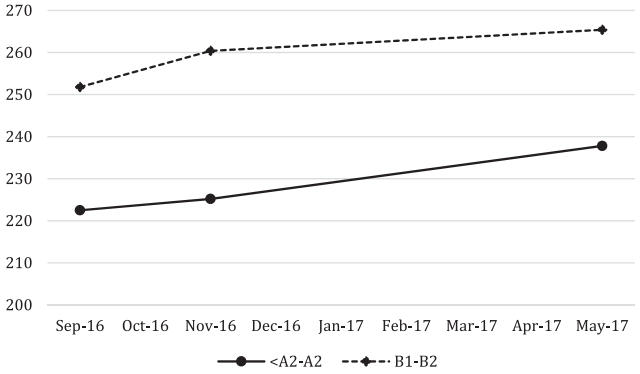

Figure 6 Students' progress in Reading scores measured at three testing times marked as reading 1, 2, and 3. The progress data are presented by proficiency level and instructional time (A1 – A2: 20 hr per term; B1 – B2: 40 hr per term).

Table 2 Means (and Standard Deviations) of TOEFL Junior Standard Across Three Test Times by Total and Section Scores (<A2 – A2 Students in the 20-hr Program, $n = 63$; B1 – B2 Students in the 40-hr Program, $n = 91$)

	<A2 – A2 students		B1 – B2 students	
Test section and time	*M*	*SD*	*M*	*SD*
Listening 1	226.3	18.9	251.6	24.0
Listening 2	236.3	21.1	269.3	20.8
Listening 3	236.3	21.6	268.2	19.9
Form and meaning 1	231.9	16.8	266.6	22.4
Form and meaning 2	237.9	18.5	269.0	18.6
Form and meaning 3	250.9	19.3	278.1	18.7
Reading 1	222.5	19.9	251.8	26.0
Reading 2	225.2	21.1	260.4	23.8
Reading 3	237.8	22.9	265.4	21.9
Total 1	680.7	44.1	770.1	62.8
Total 2	699.5	53.3	798.7	56.6
Total 3	725.1	51.7	811.8	50.1

Table 3 Differences of Means Between Three Test Times by Total and Section Scores (<A2 – A2 Students in the 20-hr Program, $n = 63$; B1 – B2 Students in the 40-hr Program, $n = 91$)

Test section	Compared period	<A2 – A2 students	B1 – B2 students
Listening	Time 2 and Time 1	10.1	17.7
Form and meaning	Time 2 and Time 1	6.03	2.36
Reading	Time 2 and Time 1	2.69	8.63
Total	Time 2 and Time 1	18.8	28.7
Listening	Time 3 and Time 1	10.1	16.6
Form and meaning	Time 3 and Time 1	19.0	11.5
Reading	Time 3 and Time 1	15.2	13.7
Total	Time 3 and Time 1	44.4	41.8
Listening	Time 3 and Time 2	0.00	−1.09
Form and meaning	Time 3 and Time 2	13.0	9.12
Reading	Time 3 and Time 2	12.5	5.05
Total	Time 3 and Time 2	25.6	13.1

DID Analysis

Below is the model for the data analysis:

$$score_i = \alpha_0 + \alpha_1\ Test2_i + \alpha_2\ Test3_i + \alpha_3\ Program_i + \alpha_4\ Test2_i\ Program_i + \alpha_5\ Test3_i\ Program_i + \epsilon_i$$ where:

α_0 = a *constant*, which shows the average score of Test 1 in the 20-hr program. It is used as a reference to interpret the remaining coefficients:

$$Test2_i = \begin{cases} 1 & \text{the score of Test 2 of the } i_{\text{th}} \text{ student} \\ 0 & \text{otherwise} \end{cases}$$

$$Test3_i = \begin{cases} 1 & \text{the score of Test 3 of the } i_{\text{th}} \text{ student} \\ 0 & \text{otherwise} \end{cases}$$

$$Program_i = \begin{cases} 1 & i_{th} \text{ student in the } 40-\text{hr program} \\ 0 & i_{th} \text{ student in the } 20-\text{hr program} \end{cases}$$

$$Test2_i Program_i = \begin{cases} 1, & \text{the score of Test 2 of the } i_{\text{th}} \text{ student in } 40-\text{hr program} \\ 0, & \text{otherwise} \end{cases}$$

$$Test3_i Program_i = \begin{cases} 1 & \text{the score of Test 3 of the } i_{\text{th}} \text{ student in } 40-\text{hr program} \\ 0 & \text{otherwise} \end{cases}$$

ϵ_i = random error of the model.

Table 4 summarizes the coefficients in this model. Specifically, a_4 is the effect of the difference in the programs in Test 2 compared to Test 1, and a_5 is the effect of the difference in the programs in Test 3 compared to Test 1.

As mentioned earlier, one challenge with this DID analysis is that the scores for Test 1 are not homogeneous across the students, so that we cannot see the effect of the instructional intensity clearly. Therefore, we propose two approaches to analyzing the data. In the first one, we analyze the data with heterogeneous initial test scores from all student participants ($N = 154$: $n = 63$ in the 20-hr program and $n = 91$ in the 40-hr program). In the second analysis, we apply the same model to a smaller subset of the data with the students whose initial scores were more homogenous, i.e., within 700 – 800 points ($N = 64$: $n = 17$ in the 20-hr program and $n = 47$ in the 40-hr program).

Analysis Based on Data With Heterogeneous Initial Scores

Table 5 shows the results of the DID analysis with all 154 student participants. Overall, the differences of all test sections and total scores are significant, with R^2 ranging from .337 to .449. Further, we summarize the results for each program separately.

Regarding the 20-hr program, the first row shows the average scores of Test 1 for this group of students (row α_0). The total score improved across all three test times. Listening improved significantly from Test 1 to Test 2, but then this growth stopped between Test 2 and 3 (rows α_1 and α_2). The form and meaning score improved significantly from Test 1 to Test 2

Table 4 Interpretation of the Coefficients

	20-hour program	40-hour program	Difference between programs	Difference between programs and tests (DID)
Test 1	α_0	$\alpha_0 + \alpha_3$	α_3	
Test 2	$\alpha_0 + \alpha_1$	$\alpha_0 + \alpha_1 + \alpha_3 + \alpha_4$	$\alpha_3 + \alpha_4$	α_4
Test 3	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_2 + \alpha_3 + \alpha_4$	$\alpha_3 + \alpha_5$	α_5

Table 5 Difference in Differences Results for the Total and Section Scores Between the 20-hr and 40-hr Programs

	Total score	Listening	Form and meaning	Reading
α_0 C	680.7***	226.3***	231.9***	222.5***
	(6.81)	(2.67)	(2.43)	(2.89)
α_1 Test 2	18.8*	10.079***	6.03*	2.69
	(9.63)	(3.78)	(3.44)	(4.09)
α_2 Test 3	44.4***	10.1***	19.1***	15.2***
	(9.63)	(3.78)	(3.44)	(4.09)
α_3 40-hr program	89.3***	25.38***	34.7***	29.2***
	(8.86)	(3.48)	(3.16)	(3.76)
α_4 Test 2: 40-hr program	9.87	7.61	−3.67	5.93
	(12.53)	(4.92)	(4.47)	(5.32)
α_5_Test 3: 40-hr program	−2.61	6.51	−7.56*	−1.56
	(12.5)	(4.91)	(4.47)	(5.32)
Observations	462	462	462	462
R^2	.449	.375	.426	.337
Adjusted R^2	.443	.368	.420	.330
Residual *SE* ($df = 456$)	54.0	21.2	19.3	22.9
F statistic ($df = 5; 456$)	74.4***	54.7***	67.8***	46.4***

Note. *$p < .1$; **$p < .05$; ***$p < .01$.

and Test 3 (rows α_1 and α_2). The reading score did not improve significantly from Test 1 to Test 2, but it did by Test 3 (rows α_1 and α_2).

Regarding the 40-hr program, the coefficients in row a_3 show the baseline difference of average scores between the two programs at Test 1. The baseline scores in this model are the scores of Test 1 in the 20-hr program. The average scores here are significantly higher for the total test score and all test sections. However, coefficients a_4 and a_5 show the poor effect of the 40-hr program compared to the 20-hr program. These coefficients are not significant for the total and section scores except for form and meaning in row α5. Moreover, this coefficient is negative with statistical significance, meaning that the effect of the program for form and meaning was reversed. We can conclude that given the scores in the 20-hr and 40-hr programs, the TOEFL Junior test is sensitive to the language gains in the 20-hr program but not in the 40-hr program.

Analysis Based on Data with Homogenous Initial Scores

One major conclusion from these results is that TOEFL Junior seems to be more sensitive to the learning gains among A1 – A2 level learners. Figure 7 visualizes this conclusion. The X axis shows the initial scores (Test 1), and the Y axis shows the change of scores from Test 1 to Test 2. It is obvious that the higher the scores in the initial test, the lower the change in Test 2.

This difference in the initial test scores makes it impossible to make more reliable estimates of the effect of intensity. To address this challenge, we analyzed a subset of our sample where the initial test scores are more homogeneous across the students in the two programs: 700 to 800 points. Figure 8 shows that in this new subset, the initial scores of the students in the 20-hr program and those in the 40-hr programs are more evenly distributed across the X axis, unlike in Figure 7. We also see that the dashed line representing the 40-hr program falls less steeply.

Table 6 summarizes the results of the DID analysis for this subset of the data. It is expected to observe fewer statistically significant differences in row α_3 because in this sample the initial scores between the students in the two programs are more homogeneous, unlike with the previous analysis (Table 5). We see some effect of the program factor on the student subsequent test scores, that is, 40 hr having a more positive impact on student scores (see rows α4 and α5). For example, we see one positive statistically significant difference for listening between Tests 1 and 3 (row α5), and the negative significant difference in the previous dataset is no longer present in these results. This trend suggests that TOEFL Junior can be sensitive to student gains in the 40-hr program too if their scores are low enough (i.e., 700 – 800 points). An important drawback in this analysis is the diluted sample size, which could have masked statistical differences where they may truly exist.

Discussion and Conclusion

At the outset, this study intended to provide more evidence for the construct validity of the TOEFL Junior Standard test to aid the stakeholders' interpretation and use of the test scores. As discussed in the introduction, Gu et al. (2015) have laid

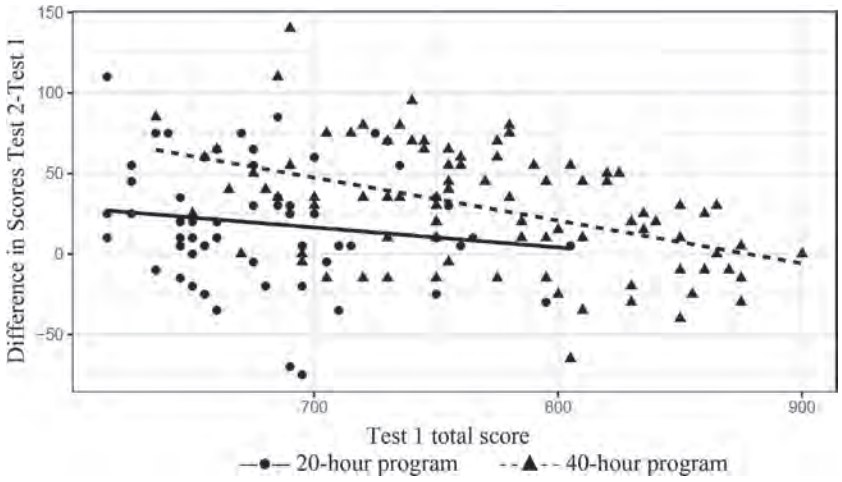

Figure 7 Relationship between Test 1 total score and the difference in scores between Tests 1 and 2 for all 154 students, by program: 20 hr and 40 hr.

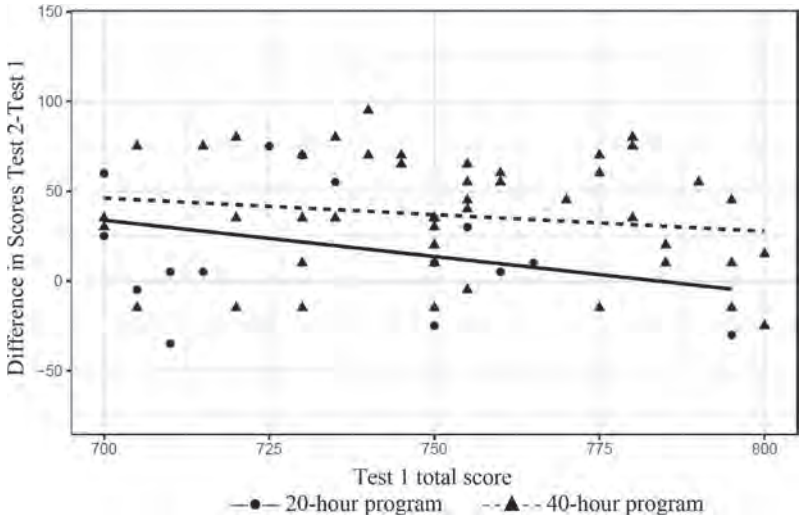

Figure 8 Relationship between Test 1 total score and the difference in scores between Tests 1 and 2 for all 64 students with homogeneous initial scores (700 – 800), by program: 20 hr and 40 hr.

Table 6 Difference in Differences Results for Total Scores and Components

	Total score	Listening	Form and meaning	Reading
α_0 C	732.6***	238.5***	251.2***	242.9***
	(8.88)	(4.19)	(3.59)	(4.70)
α_1 Test 2	20.6	14.1**	4.71	1.77
	(12.6)	(5.92)	(5.09)	(6.65)
α_2 Test 3	30.6**	9.12	10.9**	10.6
	(12.6)	(5.92)	(5.09)	(6.65)
α_3 40-hr program	20.7**	6.4	10.3**	3.97
	(10.4)	(4.89)	(4.20)	(5.49)
α_4 Test 2: 40-hr program	15.7	6.84	−0.025	8.87
	(14.6)	(6.91)	(5.94)	(7.76)
α_5 Test 3: 40-hr program	18.0	11.9*	2.84	3.24
	(14.7)	(6.91)	(5.94)	(7.76)
Observations	192	192	192	192
R^2	.299	.280	.201	.108
Adjusted R^2	.280	.260	.180	.084
Residual *SE* ($df = 186$)	36.6	17.3	14.8	19.4
F statistic ($df = 5; 186$)	15.9***	14.4***	9.38***	4.49***

Note. *$p < .1$; **$p < .05$; ***$p < .01$.

a strong foundation toward the construct validity argument of the test. However, their study was not able to account for the learning experiences of the test takers. This study adds new validity evidence by situating the study within the same after-school program, thus reducing the variability of learning experiences among the students.

Overall, the study corroborates the findings of Gu et al. (2015) that the TOEFL Junior Standard test is able to detect change in adolescent learners' language performance over time. Coincidentally, their study also looked at testing intervals similar to those found in this study: 75 days, equivalent to about 10 instructional weeks in this study and 150 days, equivalent to about 20 instructional weeks in this study. The third interval of 250 days is close to 30 instructional weeks in this study given two school breaks in this period: Christmas and spring break in the after-school program. It would not be meaningful to compare specific means of gains in the study of Gu et al. over the same periods of time with the results of this study because of unknown differences in the learning experiences of their participants. However, in general, our study supports existing evidence that learning gains within 10 weeks (20 hr of instruction) can be detected with the TOEFL Junior Standard test. Further, we discuss these findings in more detail.

This study analyzed the same periods of learning for the students with different initial scores and in the programs of different intensity. Our findings suggest that although TOEFL Junior is sensitive to learning gains as short as 20 instructional

hours over 10 weeks among A1 – A2 level students, it is not as sensitive when it comes to the learning gains of students of B1 – B2 levels of proficiency. Even though B1 – B2 level students studied for twice as many hours (40 hr in 10 weeks), this intensity does not help make the learning gains measurable. In their second model for student gains, Gu et al. (2015) included the test takers' initial scores, and as a result their estimated effects of the interval (i.e., presumed duration of study) became smaller. They suggested that "it is possible that part of the relationship between interval and gains reflects selection of how long to wait to retest based on the initial score" (p. 9). Our findings suggest that in addition to this potential influence, the initial score could be a variable that impacts student gain, hence the ability of TOEFL Junior to capture it. The literature review above cites some evidence from IELTS suggesting that language learners from higher levels of proficiency take longer to show learning gains measurable by standardized tests (Elder & O'Loughlin, 2003; Green, 2004; Green & Weir, 2003). Vocabulary research has also accumulated strong evidence showing that the lower the frequency of the word, the less it contributes to the text coverage (e.g., Nation, 2006; Nation & Waring, 1997). For example, if the first 1,000 most frequent words provide approximately 72 – 75% coverage of written text, then the second 1,000 words add only another 5% of word coverage. In addition, there is much evidence showing a strong relationship between learners' vocabulary size and their reading and listening comprehension (e.g., Milton, 2010). In light of this, it is reasonable to expect that the TOEFL Junior Standard test may not be equally sensitive to learning gain of more advanced learners compared to those of A1 – A2 learners.

Further, we would like to discuss the results for the listening section. Our statistics show that the listening scores did not grow consistently across the three testing times. There are two alternative interpretations: (a) there is some irregularity in the listening scores and (b) there is no irregularity in these scores. Regarding the first alternative, Gu et al. (2015) pointed out that one of their analyses did not show statistically significant gains for listening either. They suggest that "limited exposure to aural input in English could have hindered listening skill development" (p. 9). Even though theirs was a cross-sectional study, Wolf and Steinberg (2011) did not report a similar lack of progress in listening among their students in the United States, which again supports the argument of Gu et al. because of the ESL context. So one possible explanation is that, indeed, the gains in listening skill are harder to detect with the TOEFL Junior Standard test in non-English-speaking contexts simply because this skill is harder to develop over shorter periods of time due to the limited aural exposure. Therefore, there may simply have been no appreciable gains to be detected.

Regarding the same interpretation of the irregularity of the listening scores, the proctors ruled out the possibility of a technical problem during the last testing session. However, they reported that during the first testing session, a few younger students had left some items in the listening section unanswered because they were too tired. Perhaps listening in English does create a lot of stress, especially for younger test takers, who are also likely to have a lower level of proficiency, which may cause additional frustration. This test-taking behavior must have pulled down the mean for the first listening score, thus increasing the gain between testing Time 1 and Time 2 and decreasing this gain between testing Time 2 and Time 3.

Further, some may propose that teaching and perhaps learning in the after-school program was less effective during the winter and spring terms. However, this proposition is unlikely because the students attended multiple different classes with different teachers and different textbooks. Also, the variability of students' scores does not show any irregularities (see standard deviations in Table 2). Related to the potential seasonal influences due to the 3-week Christmas break between the fall and winter terms, there is a possibility that students' English proficiency may have decreased during this period of likely disuse. If so, this impact would obviously apply to all section scores, not only listening.

A final explanation for this irregularity is that the listening section of the second form of the TOEFL Junior Standard test was harder than it was supposed to be. This is a difficult proposition to make, but perhaps it might provide some insight to ETS test developers.

The second interpretation regarding the findings on the listening scores is that there is no irregularity. This interpretation assumes that language proficiency is not a linear function of instructional time. Under this assumption, it is possible to conclude that proficiency scores can grow and level off from time to time. Additionally, this plateau or negative difference could be due to a measurement error. Given the means and standard deviations, the standard error could easily reach 2 points, which accounts for the negative difference of −1.09 points between testing Time 2 and Time 3 among B1 – B2 level students.

Finally, it is important to highlight a few limitations in this study. Although the student participants shared learning experiences from the same after-school curriculum, they all attended their secondary schools with additional exposure to English learning practices. This must have added some unpredictable variability to the data. Next, given the curriculum

setup, it was impossible to tease apart the learners' proficiency level from the program they attended, that is, 20-hr versus 40-hr tracks. It would have been ideal to control for these two variables in the analysis. Finally, our data show that the students in the 40-hr program were 13.9 years old, and those in the 20-hr program were 12.7 years old. This age difference may have been an additional variable that impacted students' progress over time due to cognitive differences. This study could not account for this confounding variable.

In terms of delimitations, the results could be generalizable to the given after-school context with its curriculum and surrounding sociocultural and educational circumstances in the country. Gender and age in our sample were also reasonably representative of the typical population of the after-school program.

In sum, the study provides additional supporting evidence for the construct validity of the TOEFL Junior Standard test. The test shows better sensitivity to learning gains among A1 – A2 level learners than those of B1 – B2 language learners. This study also may corroborate some evidence in the literature concerning challenges with measuring listening comprehension over shorter periods of time in EFL contexts, although this conclusion should be treated with caution.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Chapelle, A. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 1 – 25). Routledge.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gains on IELTS. *IELTS Research Reports*, *4*, 207 – 254.

ETS. (2020). *Handbook for the TOEFL Junior® Standard test*. https://www.ets.org/s/toefl_junior/pdf/toefl_junior_tests_handbook.pdf

Green, A. (2004). Making the grade: Score gains on the IELTS writing test. *Research Notes*, *16*, 9 – 13.

Green, A., & Weir, C. (2003). *Monitoring score gain on the IELTS academic writing module in EAP programmes of varying duration. Phase 2 report*. UCLES.

Gu, L., Lockwood, J., & Powers, D. E. (2015). *Evaluating the* TOEFL Junior® *Standard test as a measure of progress for young English language learners* (Research Report No. RR-15-22). ETS. https://doi.org/10.1002/ets2.12064

Ling, G., Powers, D. E., & Alder, R. M. (2014). *Do* TOEFL iBT® *scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument* (Research Report No. RR-14-09). ETS. https://doi.org/10.1002/ets2.12007

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5 – 11. https://doi.org/10.3102/0013189X018002005

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211 – 232). Eurosola.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59 – 82. https://doi.org/10.3138/cmlr.63.1.59

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6 – 19). Cambridge University Press.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). TOEFL Junior® *design framework* (Research Report No. RR-15-13). ETS. https://doi.org/10.1002/ets2.12058

Wolf, M. K., & Steinberg, J. (2011). *An examination of United States middle school students' performance on TOEFL* Junior® (Research Memorandum No. RM-11-15). ETS

Suggested citation:

Madyarov, I., Movsisyan, V., Madoyan, H., Galikyan, I., & Gasparyan, R. (2021). *New validity evidence on the* TOEFL Junior® *Standard test as a measure of progress* (TOEFL Research Report No. RR-95). ETS. https://doi.org/10.1002/ets2.12334

Action Editor: John Norris

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