Listening. Learning. Leading.®

Research Report
ETS RR–15-22

Evaluating the *TOEFL Junior*® Standard Test as a Measure of Progress for Young English Language Learners

Lin Gu

John Lockwood

Donald E. Powers

December 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluating the *TOEFL Junior*® Standard Test as a Measure of Progress for Young English Language Learners

Lin Gu, John Lockwood, & Donald E. Powers

Educational Testing Service, Princeton, NJ

Standardized tests are often designed to provide only a snapshot of test takers' knowledge, skills, or abilities at a single point in time. Sometimes, however, they are expected to serve more demanding functions, one of them is assessing change in knowledge, skills, or ability over time because of learning effects. The latter is the case for the newly developed *TOEFL Junior*® Standard test, which measures improvement in young learners' proficiency in English as a foreign language. In this study, we used nonexperimental-repeated measures data from approximately 4,600 students from multiple countries to examine the extent to which observed patterns in within-individual changes in test scores were consistent with changes in underlying language proficiency because of learning. Because most students were actively participating in English language learning programs, the time interval between test administrations, which varied among students, served as a proxy for the extent of English language learning opportunities. We used hierarchical linear models to model growth in test performance as a function of the time interval between test administrations and found a positive, statistically significant relationship; that is, test takers with longer intervals between retesting exhibited greater gains than did test takers who retested at shorter intervals. The estimated relationship for the total score corresponded to between .16 and .24 test standard deviations of growth per year, depending on model specification. The findings are robust to sensitivity analyses that explore potential biasing factors. Overall, the findings are consistent with the hypothesis that the TOEFL Junior Standard test is capable of reflecting change in English language proficiency over time.

Keywords young learners; English language proficiency; growth; longitudinal; hierarchical linear model; TOEFL Junior® Standard test

doi:10.1002/ets2.12064

Often, standardized tests are designed primarily to provide only a snapshot of test takers' knowledge, skills, or abilities at a single point in time. This is typically the case when test scores are used for selection, admission, or certification purposes, for example. Some tests, however, are expected to meet more demanding functions, one of them is assessing change in knowledge, skills, or ability over time. The measurement of change over time is an important area in educational research because it offers a means to evaluate the effectiveness of educational efforts, which are intended generally to effect changes in students' attitudes, achievement, and values (Carver, 1974; Willett, 1994). Currently, assessing change (or growth) is especially important in one domain in particular in which millions of young learners are engaged worldwide — English language proficiency.

Regardless of its type or purpose, every test is required to meet a variety of professional standards, including those for the interpretation and use (i.e., validity) of test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Although validity theory has been explicated in a variety of ways, most modern theorists have, in some form, advocated (a) making explicit the claims that are made for a test, (b) developing a validity argument that supports these claims, and (c) evaluating the coherence of the argument, including the plausibility of the assumptions on which it rests and the inferences that follow from it (see, e.g., Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Kane, 2006; Messick, 1989).

To argue that a test is capable of measuring change requires evidence that links changes in observed test scores to actual (or at least presumed) changes in the abilities that a test is intended to measure. Because changes in abilities often result from relevant learning experiences, evidence indicating an association between test performance and construct-relevant learning experiences can be used to support a claim that a test is sensitive to change.

Corresponding author: L. Gu, E-mail: LGu001@ets.org

Furthermore, establishing a link between test performance and relevant learning experiences also contributes to the overarching goal of construct validation; that is, determining the degree to which a test measures the target construct. For instance, both Messick (1989) and Chapelle et al. (2008) proposed that construct interpretation is facilitated to the extent that test performance changes are commensurate with the amount and quality of relevant learning experiences.

Methodologically, validity evidence can be collected either cross-sectionally or longitudinally. Several studies in language testing have used cross-sectional designs to investigate the associations between test performance and construct-relevant learning experiences. These studies have compared groups of students that differ on a variable of interest (learning experience, for example) at a single point in time. For example, test performance in various testing situations has been examined in relation to learning experience, such as out-of-school contact with French (Morgan & Mazzeo, 1988) or Spanish (Ginther & Stevens, 1998) as the target language, experience with Korean in an at-home heritage environment (Bae & Bachman, 1998), exposure to English as a foreign language in either educational or business contexts (Stricker & Rock, 2008), and English-as-a-second-language study-abroad environments (Gu, 2014). However, to our knowledge, virtually no studies in the field of language testing have used a longitudinal design to undertake such validation investigations with the goal of evaluating the appropriateness of score interpretation and use. One exception is a study by Ling, Powers, and Alder (2014), which used a pretest and posttest design to investigate the extent to which the *TOEFL iBT*® practice test reflects the learning effects in different English-language programs for adult English language learners.

This study uses a longitudinal design to model changes in relation to learning over time. The goal was to examine the extent to which a new assessment of young learners' English language ability, the TOEFL Junior® Standard test developed by Educational Testing Service, is capable of reflecting changes in language ability as a function of learning. This property (i.e., sensitivity to changes because of education) is especially crucial for tests designed for learners of young ages. Young learners' abilities develop as a result of their engagement in learning in various contexts, both within and outside of the school environment. To be maximally useful therefore, an assessment developed for young learners should exhibit an ability to reflect growth that accrues from construct-relevant learning.

The TOEFL Junior Standard Test

The TOEFL Junior Standard test is a test designed for adolescent English language learners (primarily between 11- and 15-years old). It measures English language proficiency with respect to the academic and social English language skills needed to meet the demands that young English language learners face in English-medium instructional environments. Test development is based on expectations for middle school students in English-medium secondary schools as informed by a variety of sources. Test tasks are based on both social and academic uses of language in a school context.

The test is composed of multiple-choice questions in three sections. The listening comprehension section (referred to as the "listening" section here for brevity) is designed to measure the following key skills: understanding main ideas, identifying important details, making inferences based on a speaker's intonation or stress, understanding idiomatic language, and understanding how information is being used by a speaker. The language form and meaning ("language") section tests a student's ability in key English language skills such as recognizing the accurate meaning and use of more advanced grammatical structure (e.g., relative clause), demonstrating knowledge of a wide range of vocabulary that includes words found primarily in academic texts, and recognizing how sentences combine to create cohesive, meaningful paragraphs. The reading comprehension ("reading") section measures a student's ability to understand main ideas, comprehend important details, make inferences, infer the attitude or point of view of a character in a fictional story, understand figurative language, and determine the meaning of unfamiliar vocabulary words from contexts.

One of the major proposed uses of the TOEFL Junior Standard test is to provide objective information about student progress in developing English language skills over time. Language learning usually occurs in two contexts — an instructional context and a communicative context (Batstone, 2002). In an instructional context, learners often develop their language skills from instruction received in a classroom setting where English is taught as subject matter. Formal foreign language training in the home country usually provides such a context. In a communicative context, the objective is to use the target language to perform communicative functions. Study abroad in the target language community is likely to create such communicative contexts for learners. If used repeatedly, the TOEFL Junior Standard test is expected to be able to reflect anticipated gains in the target construct that result from learning in various contexts.

Research Objectives

Our primary research goal was to examine the degree to which observed patterns in within-individual changes in the TOEFL Junior Standard test scores are consistent with expected changes in underlying language ability because of learning. The challenge here was to disentangle several possible sources of within-individual changes in test scores, including true learning, test familiarity/practice, and simple random measurement error. The challenge was heightened because no direct measures of instructional experiences existed for the students in our study sample. However, we did know that all students were engaged in instructional programs intended to improve English skills, both in school and perhaps outside of school as well. Therefore, the time interval between test administrations served as a proxy for the amount of English language learning opportunities. The specific research question investigated was this: to what extent do the scores of the TOEFL Junior Standard test reflect changes in language ability as a function of time elapsed between repeated observations? Our analysis thus considers whether larger score gains were exhibited by students with longer intervals between test administrations. The analysis rests on the assumption that, because of true learning, scores should increase as a function of the time interval between administrations. Therefore, a significantly positive relationship between interval and score gains would provide at least circumstantial evidence in support of the claim that the test is capable of reflecting changes in English language ability over time.

Method

Data

As mentioned previously, the study used a longitudinal data collection design. This design was made possible by the test developer's communication to test users that the test could be used to monitor growth in English language learning. As such, a sufficient number of schools and other educational units worldwide have been using the test to monitor student progress, giving rise to longitudinal data on individual students.

Data were retrieved from countries where the test was being administered between early 2012 and mid 2013. Beyond simply encouraging test users to retest students at approximately 6-month intervals, no control was imposed on the frequency of retesting, and indeed, many students took the test only once. Students who took the test multiple times were tested various numbers of times and at variable intervals between test administrations, in most cases according to schedules set by local education agencies or individual schools. Each data record contained a unique student identifier that permitted tracking of students across multiple test administrations.

The majority of the test takers in the base dataset took the test only once. The rest took the test more than once. Because we are interested in how student test performance changes over time, our analysis focuses only on students who took the test more than once. On the initial administration, these repeat test takers scored only slightly lower (about .02 – .03 test standard deviation units) than those who tested only once. This suggests that while repeat test takers were a relatively small fraction of the total examinee population, they were initially not very different in terms of English proficiency.

Among the repeat test takers, 4,606 students had complete test data and were therefore included in the analysis sample for the study. A total of 15 countries and regions were represented in the sample. More than half of the test takers (about 65%) were from Korea. Table 1 summarizes the distribution of number of test administrations in the analysis sample. As shown in the table, the vast majority of repeat test takers ($N = 4{,}205$) took the test exactly twice.

The data had a hierarchical structure. At the highest level, students were nested in countries.[1] Within each country, the grouping structure was more complicated. Students were members of various groups that could influence their test scores, but we had imperfect information about those groupings. For example, our data did not have links of students to schools, teachers, or other educational programs. We also knew that groups of students were tested together at testing centers, but we did not have test center information in our data, so it was impossible for us to definitively identify groups of students who took the test at the same place at the same time. The grouping structures of the data are relevant to building statistical models that account for potential sources of variances in the test scores. Therefore, we needed to approximate the true grouping structures as accurately as possible given the available data.

The closest proxy we had to schooling experiences for individual students was the so-called client. Often, the client corresponded to an instructional provider (e.g., a school district, a school, or a learning center) and so could be thought of as proxy for links of students to schools. The client in many cases was responsible for arranging for the testing of a group of students on certain dates. Therefore, by identifying a group of repeat test takers who were linked to the same

Table 1 Summary of Distribution of Number of Test Administrations in the Repeater Analysis Sample

Number of times test taken	Number of students	Percentage of repeater sample
2	4,205	91.3
3	252	5.5
4	72	1.6
5	75	1.6
6	1	0
7	1	0
All repeaters	4,606	100

client and who took the tests on the exact same dates, we could be reasonably confident that those students were likely to have shared experiences, including similar instruction and similar testing conditions that could be related to test scores. We refer to such a group of students as a testing group. In our analyses, we used testing groups as a grouping variable nested within countries. Students sharing a testing group had, by definition, the same amount of time elapsed between repeat test administrations. Across testing groups, testing schedules (and therefore the time elapsed between administrations) varied, providing the key variation that allowed us to examine the relationship between test scores changes and time between administrations.

Testing groups were intended to serve as a reasonable, but not perfect, proxy for relevant grouping structures. For example, not all clients corresponded to educational entities because individuals could choose to take the exam for reasons unrelated to group educational experiences, such as applying to some educational program that requires demonstrated English proficiency. In our data, 7.8% of the repeat test takers did not appear to be part of any testing group because the specific dates on which they were tested did not match other students in the data. We included these students in the analysis and assigned each one to a unique testing group. We also tested the sensitivity of our findings to the exclusion of these students.

Descriptive Analysis

For our analysis, we used scaled scores for each of the three test sections (language, listening, and reading) as well as the total score based on the three subscores. Scores ranged from 200 to 300 for each section and from 600 to 900 for the entire test. Different forms of the test are rigorously equated to ensure that scores from alternate forms are comparable, and in the appendix, we present sensitivity analyses indicating that our findings were not sensitive to the use of different forms.

We first grouped students by the total number of times they took the test and examined their average scores over the course of the repeated administrations. Figure 1 defines groups of students by how many times in total they took the test and then plots the average total scale score (vertical axis) against the average amount of time (in days) between administrations (horizontal axis). We use the term interval to refer to the number of calendar days between administrations. Two students who took the test more than five times were excluded from the plot. The figure demonstrates that, on average, students' scores increased over the course of the repeated administrations. The different groups generally showed the same increasing trend upon retesting. Trends for the language, listening, and reading subscales (not shown) were similar. Given that more than 90% of the analysis sample took the test only twice, we focused the remainder of our analyses on only the first two test scores from all students in the analysis sample, regardless of how many times they took the test.

Figure 2 provides a histogram of interval between the first and second administrations for the analysis sample ($N = 4{,}606$). The distribution is multimodal with the largest mode centered around 180 days. According to the *Handbook for the TOEFL Junior Standard Test* (Educational Testing Service, n.d.), an interval of about 6 months between repeated test administrations is suggested for students taking a regular English curriculum to show gains in their scores. The largest mode in Figure 2 confirms that most students were retested at approximately 6-month intervals but substantial variation is present around this mode.

Based on the distribution of time interval between testing and retesting, we decided to separate the students into four groups for descriptive purposes. Students in the first group ($N = 619$), the shortest interval group, had less than or equal to 75 days (about 2.5 months) between testing and retesting. Students in the second group ($N = 467$) had an interval between 75 and 150 days (approximately 2.5–5 months). Students in the third group ($N = 2{,}738$) had an interval between 150

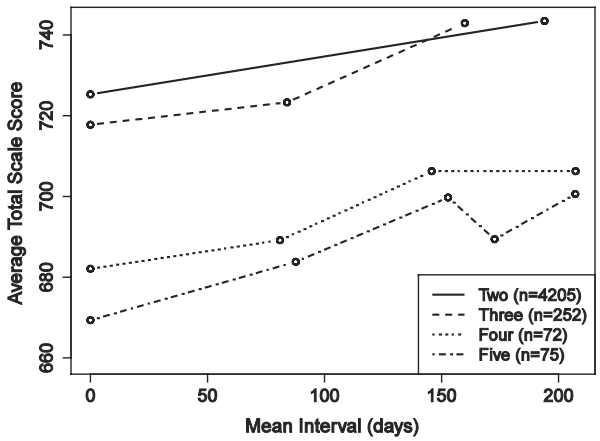

Figure 1 Average total scale score by test administration for students who took the test two, three, four, or five total times. Scores for repeat administrations are plotted by the average interval (in days) between administrations for each group of students.

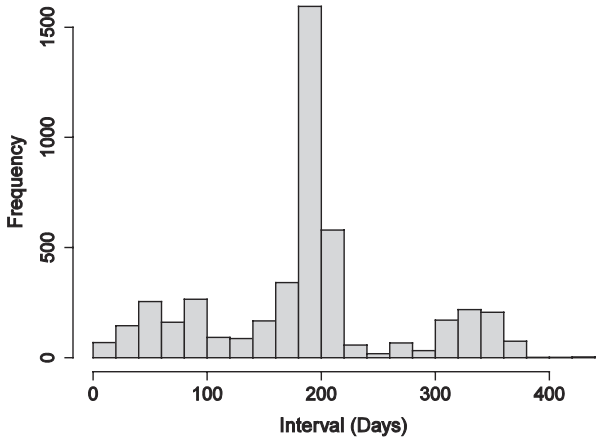

Figure 2 Histogram of interval for the analysis sample.

and 250 days (approximately 5 – 8.3 months), and the longest interval group ($N = 782$) had more than 250 days between retesting. The cut points for these intervals approximately correspond to points that would separate the different modes evident in Figure 2. These groups were used only for the descriptive purposes presented now, and all further analyses we present use the continuous values of interval measured in days between administrations.

Table 2 provides the average scores for both test administrations and the average gains for each of the three groups defined by interval between administrations. The table indicates that score gains generally increased as the intervals lengthened for both the subscales and the total score. The longest interval group had the largest score gains, while the shortest interval group had the smallest score gains.

An alternative display of the relation of average gains to between-test interval is provided in Figure 3, where interval is treated as a continuous variable. Figure 3 presents estimated smoothed conditional mean of the first and second administration scores as a function of interval. The horizontal axis in each frame of the plot is the interval between testings, and the vertical axis is scaled test score. A separate frame is shown for each section score. The solid curve in each frame is the estimated mean score on the first test administration as a function of interval, while the dashed curve is the estimated mean score on the second test administration as a function of interval.

As shown in Figure 3, the distance between the dashed and solid curves, representing average score gains, tends to increase as the interval increases. Generally, the effect is modestly evident from the low end of the interval distribution to the middle and very evident at the higher end of the interval distribution, consistent with Table 2, in which the interval was treated as a categorical variable. The two curves start to diverge at an interval of about 200 days, and this pattern is consistent for all subtests.

Table 2 Mean First and Second Scores and Mean Gain for Each Outcome by Interval

Outcome	Interval	First	Second	Difference
Language	Interval ≤ 75	241.1	243.5	2.4
	75 < Interval ≤ 150	241.5	245.4	3.9
	150 < Interval ≤ 250	240.1	243.3	3.2
	Interval > 250	242.1	255.3	13.2
Listening	Interval ≤ 75	245.1	247.6	2.5
	75 < Interval ≤ 150	247.8	250.5	2.7
	150 < Interval ≤ 250	246.8	251.1	4.3
	Interval > 250	242.6	253.0	10.4
Reading	Interval ≤ 75	237.4	240.9	3.5
	75 < Interval ≤ 150	236.2	242.1	5.9
	150 < Interval ≤ 250	234.7	242.2	7.5
	Interval > 250	242.7	253.9	11.2
Total	Interval ≤ 75	723.5	732.0	8.4
	75 < Interval ≤ 150	725.5	738.0	12.5
	150 < Interval ≤ 250	721.6	736.6	15.0
	Interval > 250	727.4	762.2	34.8

We noticed, however, that the higher end of the interval distribution is dominated (about 70%) by students from a single country that otherwise had a much smaller percentage of test takers in the sample. We found out that these students were enrolled in a 2-year learning program that used the test either yearly or bi-yearly. This explained why the average interval was much longer for this group. Although these students may constitute a distinct population, they do not differ much from students in other countries with respect to overall initial score.

Hierarchical Linear Models

The descriptive information is consistent with the hypothesis that the test is sensitive to changes in English ability as a function of interval. We also conducted analyses that used hierarchical linear models (HLMs; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002) to estimate the relationship between interval and student gains. The main motivation for conducting HLM analyses was to quantify the relationship between interval and gains in a way that would not be distorted by potential differences in student populations across countries and to obtain a valid test of statistical significance of this relationship. Our HLM is designed to achieve both of these goals. We first present what we call the base model and then present what we call the alternative model, which we use to address a particular potential source of bias in the base model.

Base Model for Student Gains

Our base model for student gains is

$$\left(Y_{i2} - Y_{i1}\right) = \mu_{j(i)} + \beta\, INTERVAL_i + \theta_{g(i)} + \varepsilon_i \qquad \text{(Model 1)}$$

where $(Y_{i2} - Y_{i1})$ is the gain score for student i on whatever outcome is being modeled (language, listening, reading, or total scores); $\mu_{j(i)}$ is a dummy variable for the country j in which each student i is nested; $\theta_{g(i)}$ is a random effect for the testing group g for student i is assumed to be mean zero, normally distributed, and independent across testing groups g with common variance; and ε_i is a residual error assumed to be mean zero, normally distributed, and independent across students with common variance. *INTERVAL* refers to the number of calendar days between the first and second administrations. The effect of interest is the coefficient on interval.

All other effects in the model are included only to improve the estimate of the effect of interval and to get an appropriate measure of uncertainty about this effect. We now describe these other effects and why they should achieve these goals.

We include dummy variables in the model for individual countries to prevent country differences from biasing our estimated relationship between interval and gains. As noted previously, there were large differences in the distributions of interval across countries, especially at the higher end of the interval distribution. These differences could have led to bias in the estimated relationship between interval and score gains if students from different countries had systematically

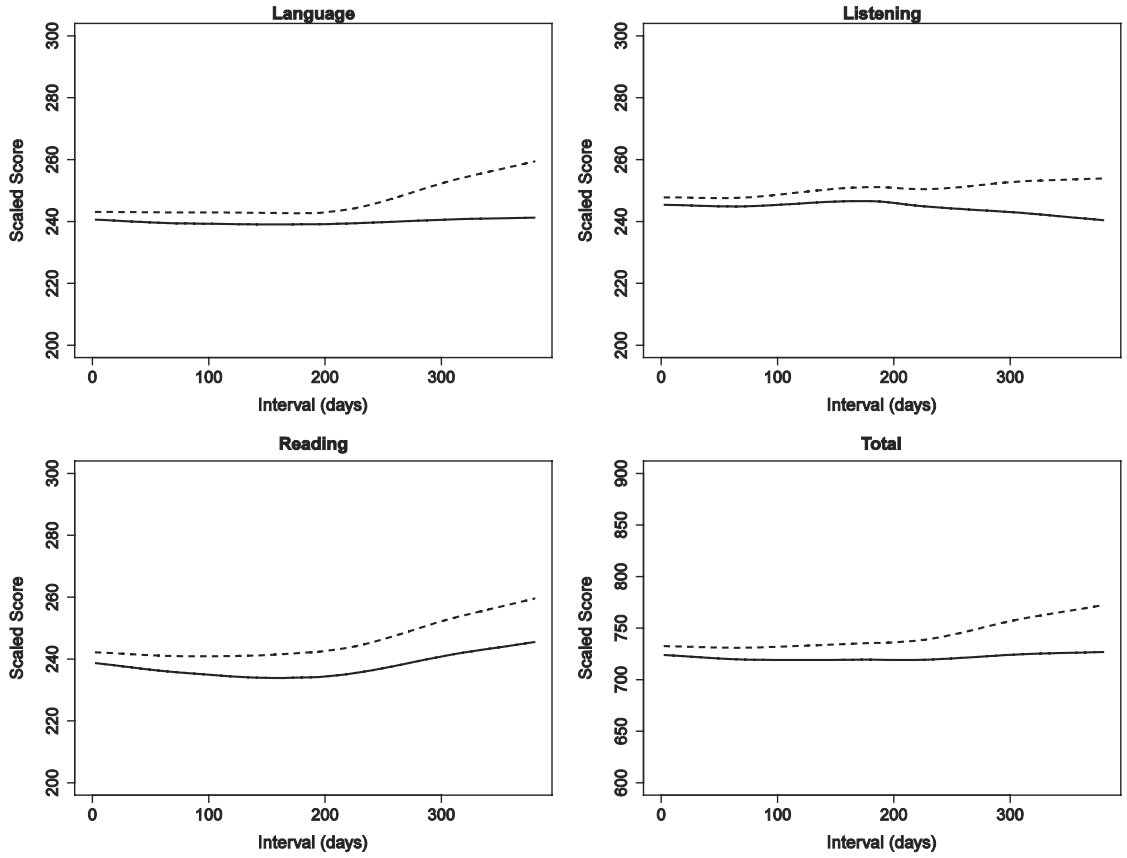

Figure 3 Smoothed trends of initial score (solid line) and second administration score (dashed line) as a function of interval, separately for each skill area and the total.

different average values of interval because of unobserved test taker characteristics that were related to score gains. It seemed reasonable to assume that such characteristics might exist, so we needed to treat differences among countries as a potential confounding factor and find a way to prevent this potential confounding factor from biasing the estimated coefficient on interval. Treating countries as dummy variables avoids this potential confounding, as it ensures that the effect of interval is estimated using only variation in interval among students but within countries (i.e., by comparing gains for students in the same country but with different intervals). Differences in the average value of interval across countries do not therefore contribute to the estimated effect of interval, and therefore, differences among students in different countries cannot be a source of bias.

We include random effects in the model for testing groups. By definition, students in the same testing group had the same interval; thus, it was not possible to include dummy variables for testing groups because that would have left no variation with which to estimate the effect of interval. However, it was possible to include random effects for testing groups. Doing so is beneficial and important. As described earlier, students belonging to the same testing group were likely to share similar instructional or testing experiences. These factors could have caused students who shared a testing group to have systematically higher or lower gains than would have been predicted based on only the interval and the country for that group. This is a form of clustering in the residual for the model, which, if not accounted, could lead to an inefficient estimate of the effect of interval as well as an overstatement of the precision of the estimated effect (Raudenbush & Bryk, 2002; Wooldridge, 2002). The testing group random effects are introduced to account for unobserved differences in the average gains in different testing groups that remain after accounting for both interval and country effects. Accounting for this variation through the random effects leads to a more accurate estimate of the effect of interval and an accurate standard error for this estimated effect.

Our model assumes a linear relationship between interval and gains within countries. It is reasonable to question whether a nonlinear relationship would be more appropriate. We tested our model against more complicated alternatives

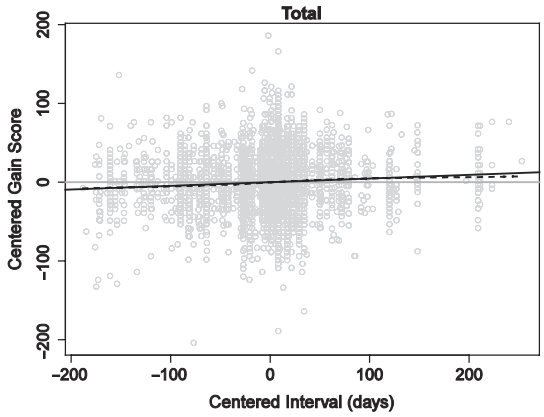

Figure 4 Scatterplot of total score gain versus interval where both are centered around their respective country means. Black-solid line is the best linear fit, and the black-dashed line is a smooth regression function allowing for nonlinearity. The horizontal gray line at 0 is provided for reference.

that allowed the relationship to be nonlinear, and the linear specification was preferred by standard model comparison criteria including a likelihood ratio test, the Akaike information criterion, and the Bayesian information criterion (Pinheiro & Bates, 2000). The linear specification is also supported by a graphical analysis. Figure 4 provides a scatterplot of total gains versus interval where both quantities are centered by their respective country means. This represents the relationship between total gains and interval within countries, aligning closely to how Model 1 identifies the effect of interval through the use of country dummy variables. The black line is the linear fit, and the black-dashed curve is a nonparametric smooth regression flexible enough to capture nonlinearity if it existed. The fact that these two curves nearly coincide supports the assumption that the relationship between gains and interval within country is well approximated by the linear specification of Model 1.

Our model also assumes a common linear relationship across countries. It is reasonable to question whether allowing the slope relating interval to gains varies across countries. We tested such a model against our simpler alternative, and again, our model was preferred by standard model comparison criteria.

Alternative Model for Student Gains

A major threat to validity of the analysis is that the data are purely observational (nonexperimental) in nature. That is, there was no experimental manipulation of the primary independent variable, interval between testing. An experiment to understand the effect of interval on outcomes would randomly assign students to different values of interval, ensuring that interval on average is unrelated to any student characteristics. In this case, a relationship between interval and outcomes would be interpretable as a causal effect of interval on outcomes, where presumably the causal mechanism is the English language learning that is taking place during the interval. As mentioned previously, in our data, interval was not experimentally controlled. Furthermore, we had only limited information about how interval was determined. For students with observed testing schedules that did not suggest obvious membership in a group testing situation, we could not discount the possibility that interval was in part determined by initial scores or by other student attributes that may have been related to outcomes. More generally, even for testing groups that had a shared testing schedule, we did not know the basis for the schedule. Although it is reasonable to assume that the interval between tests was often chosen by the educational entity providing instruction, we do not know what information was used in that decision. For example, we do not know whether the decision about when to retest was made prior to any testing or whether it was in part based on students' initial performance. The analyses using gain scores are unaffected by any relationship between interval and unobserved student characteristics that are constant across time and related to scores because the differencing used to calculate gains would negate the impact of those characteristics. On the other hand, if interval is in any way influenced by the observed initial scores of either individual students or groups of students, an analysis based on Model 1 is potentially biased. In this case, a better option would be to augment Model 1 with additional covariates of the initial total scores of both the individual students and their testing groups. The use of total scores as covariates reflects the likelihood that if decisions about when to do follow-up testing are influenced by initial scores, those decisions are most likely to be based

Table 3 Summary of Results of Model 1 of the Effect of Interval on Gains

Outcome	Estimate	*SE*	*t*-stat.	*p*-value
Language gain	.013	.005	2.76	.006
Listening gain	.010	.006	1.77	.076
Reading gain	.022	.007	3.23	.001
Total gain	.045	.012	3.69	<.001

on overall performance. The alternative model, Model 2, is

$$\left(Y_{i2} - Y_{i1}\right) = \mu_{j(i)} + \beta INTERVAL_i + \gamma X_{i1} + \delta \overline{X}_{g(i)1} + \theta_{g(i)} + \epsilon_i \qquad \text{(Model 2)}$$

where X_{i1} is the first administration total scale score for student i, and $\overline{X}_{g(i)1}$ is the average first administration total scale score for students in testing group g. All other terms are defined analogously to Model 1.

We fit both Models 1 and 2 using the routine lmer in R (R Development Core Team, 2007). The two models were fit separately to each of the three subscores and the total score.

Results

Table 3 presents estimates of the coefficient on interval for the different outcomes using the base model (Model 1). The coefficients are for interval scaled in days, so that a coefficient of, for example, .03 represents an average score gain of three points per 100 days of interval or about 11 points for a year. The estimated coefficient is positive for all outcomes and is statistically significant at the .05 level[2] for all outcomes except listening. Because most of our repeat test takers were located in a foreign language environment, we speculate that limited exposure to aural input in English could have hindered listening skill development. The estimated relationship of interval to the total score gain is .045 points per day or about 16.4 points per year. The standard deviation of the total score in the analysis sample in the second administration is about 69 points. Thus, the estimated effect corresponds to an increase of .24 standard deviation units over a 1-year interval. This magnitude of change over 1 year is consistent with findings on annual growth on numerous standardized reading exams used in the United States. As reported by Bloom, Hill, Black, & Lipsey, 2008 (Table 1), annual growth over Grades 6–9, which roughly corresponds to the median age of students in our sample of 13 years, tends to be about .24 standard deviations per year. We stipulate that because the average interval in our data is much less than a full year, these extrapolations to annual effects rely heavily on our linear model specification and should be interpreted cautiously. We also stipulate that the comparison with similarly aged English-speaking students in the United States is limited because of the obvious differences between those students and the students in our sample.

The results in Table 3 are robust to a number of sensitivity analyses involving decisions about the model and analysis sample. The details on each sensitivity analysis are included in *Appendix*. The results for the total gain are presented in Table A1. Each row is a different sensitivity analysis, with the base model results presented in the first row for ease of comparison. The conclusion from these analyses is that the positive relationship between interval and score gains is largely insensitive to changes in the sample or model for gains based on characteristics that could have affected the results.

Results of the alternative model (Model 2) are summarized in Table 4, which is analogous to Table 3. The results are similar to those from Model 1, although there is some evidence of the estimated effects being smaller. For the total score, the estimated coefficient on interval is reduced from .045 in Model 1 to .03 in Model 2, corresponding to a standardized effect size of .16 standard deviation units. The reduction in the estimated effects suggests 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. The results of this model, represented in Table A2, are as similarly robust to sensitivity analysis as the results of Model 1 described in *Appendix*.

Discussion

The primary objective of the study was to evaluate the extent to which longitudinal test data support the claim that the TOEFL Junior Standard test can serve to measure young learners' progress in learning English as a foreign language. We found that, on average, repeat test takers scored higher on the second administration and that the longer the interval

Table 4 Summary of Results of Model 2 of the Effect of Interval on Gains

Outcome	Estimate	*SE*	*t*-stat.	*p*-value
Language gain	.012	.005	2.41	.016
Listening gain	.006	.006	.95	.343
Reading gain	.013	.007	1.90	.057
Total gain	.030	.012	2.39	.017

between testing was, the greater the score gain was. The estimated relationship ranged from .16 to .24 standard deviation units of growth per year depending on the model specification. These values are consistent with annual reading achievement growth rates for similarly aged students in the United States.

We considered the following three plausible explanations for our findings: (a) observed increases were indicative of improved ability in English resulting from learning, (b) increases resulted simply from greater familiarity with the test as a result of having taken it previously (retesting effects), and (c) the relationships were because of inadequacies of our approach of using a nonexperimental proxy for English learning opportunities.

With respect to retesting effects, it seems implausible that the observed pattern of increases was due primarily to test takers having gained familiarity with the test by having taken it previously. Test practice effects are typically observed more often for tests that use complex directions and item formats (see, e.g., Kulik, Kulik, & Bangert, 1984; Powers, 1986). Neither of these qualities are characteristic of the TOEFL Junior Standard test. Moreover, because the effects of becoming familiar with a test as a result of having taken it in the past are likely to decrease over time, the longer the interval between retesting is, the less the impact on score increase is likely to be. In our analysis, we found a positive relationship between increase in test score and length of time between retesting, which is inconsistent with score increases being due simply to having taken the test previously.

Two distinct threats arise from our use of a nonexperimental proxy (interval) for English language learning. The first is that because interval was not experimentally assigned, we cannot rule out that the observed relationship between interval and gains is spurious. To the extent that any spurious relationship is driven by students with lower initial scores waiting longer to be retested, Model 2 should be effective in mitigating the bias. However, other forms of spurious relationship are possible, and those would present a source of bias in our findings. The fact that the estimated effects in Model 2 are generally smaller than those in Model 1 leaves open the possibility that such biases might exist. While we can never rule out such scenarios, we can at least be confident that the results of testing both models provided convincing evidence in support of the relationship between interval and gains and that our results are robust to a wide array of sensitivity analyses that were conducted to address potential confounding factors within the limits of the available data.

The second threat arising from our use of interval as the proxy would exist even if interval were randomly assigned and cannot easily be tested given our data. We have assumed that because students in the sample are generally participating in English instruction, interval can be treated as a proxy for English learning opportunities. However, interval serves as a proxy for all maturation processes, not just English learning. We cannot therefore rule out the possibility that gains are due, in addition to language learning, to growth in cognitive or behavioral attributes (e.g., ability to concentrate). Interval may also proxy for test preparation occurring outside the context of the formal test administrations. While we have no direct evidence that the students in our sample were engaging in test preparation, opportunities for that do exist given that the test is part of the *TOEFL*® family of assessments and that some tests in the TOEFL family have a high-stakes nature. However, the TOEFL Junior Standard test for the students in our sample has only low-to-medium-stakes implications for either individuals or programs, so the possibility that the results are due solely to narrow test practice seems unlikely. The low-to-medium stakes of the test might also explain why our estimate of the annual growth is perhaps smaller than some people might expect. It is likely that in our sample, the test was primarily being used for routine monitoring of student performance in education settings rather than for any high-stakes decisions about individual students, and so it is possible that observed growth may be larger in settings where students were highly motivated to improve performance. In any case, without tying growth in test scores directly to quality and quantity of English and instruction and demonstrating that more and better instruction produces larger gains, our evidence remains indirect.

However, the simplest explanation for our findings is that observed test score increases are due, at least in part, to real changes in the target ability as a result of English language learning. The magnitude of score increases is related in anticipated ways to the length of interval between retesting, which could be reasonably considered as a proxy for English

language learning. We acknowledge the limitation of using this proxy to represent the amount of learning undertaken by participating students over the course of repeated test administrations. In order to fully capture the richness and complexity of learning, future studies should collect test-taker background information (e.g., years of studying English, the starting age of learning, the type of instruction, the type of curricula, language use inside and outside of classroom, language exposure, and motivation) and model these as additional covariates in the analysis. Such information could also allow more accurate accounting of shared experiences of groups of students, which would improve the modeling. What also needs to be acknowledged is that slightly more than half of the study sample came from a single country. This factor may limit the extent to which the study results can be generalized to the entire target test-taking population, that is, young English language learners worldwide.

Despite the aforementioned study limitations, we believe that this study constitutes an initial step in providing evidence that the TOEFL Junior Standard test can reflect changes in language ability that result from learning. The findings therefore provide initial support for the claim that the test can be used to monitor growth for young English language learners. Finally, the findings also lay open the possibility that score reports can be enhanced by incorporating information on score change in order to provide a historical account for test takers who take the test multiple times.

Notes

1 Each student in the sample was associated with a scoring code that indicated the student's country of origin. Except for a very few countries which were associated with multiple scoring codes, most of the countries were associated with only one scoring code. We therefore decided to use the label "country" to refer to this level of nesting in the article.

2 Across Tables 3 and 4, we conducted eight hypothesis tests, five of them are significant at level .05. To address concerns about multiple testing, we also applied the Benjamini–Hochberg procedure as recommended by the What Works Clearinghouse (2014) using a false discovery rate of .05. All five of the originally statistically significant findings are still statistically significant after applying this procedure.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing*, *15*, 380–414.

Batstone, R. (2002). Contexts of engagement: A discourse perspective on "intake" and "pushed output." *System*, *30*, 1–14.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions* (MDRC Working Papers on Research Methodology). Retrieved from http://www.mdrc.org/sites/default/files/full_473.pdf

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, *29*, 512–518.

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). New York, NY: Routledge.

Educational Testing Service. (n.d.). *Handbook for the TOEFL Junior Standard Test*. Retrieved from http://www.ets.org/s/toefl_junior/pdf/toefl_junior_student_handbook.pdf

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 169–194). Mahwah, NJ: Erlbaum.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, *31*, 111–133.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education & Praeger.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement scores. *American Educational Research Journal*, *21*, 435–447.

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). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12007

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 – 103). New York, NY: American Council on Education & Macmillan.

Morgan, R., & Mazzeo, J. (1988). *A comparison of the structural relationships among reading, listening, writing, and speaking components of the AP French Language Examination for AP candidates and college students* (Research Report No. RR-88-59). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00315.x

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects. *Psychological Bulletin*, *100*, 67 – 77.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-Based Test across subgroups* (TOEFL iBT Research Report 07). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02152.x

What Works Clearinghouse. (2014). *Procedures and standards handbook version 3.0*. Washington, DC: U.S. Department of Education.

Willett, J. B. (1994). Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671 – 678). Oxford, England: Pergamon Press.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Appendix

Sensitivity Analyses

Table A1 presents results of sensitivity analyses using Model 1. It presents estimates only for the total score gain to keep the presentation of results manageable. The findings for the other outcomes are similarly robust, again with the listening subscale tending to have positive but statistically insignificant findings. Row 1 repeats the base model of Table 3 for comparison. Row 2 adds to the model a dummy variable for each unique pair of test forms that the students took on their first and second administrations to test for the possibility that form-to-form variation in either content or scaling is potentially confounding the results. Row 3 removes all students with intervals less than 15 days or greater than 400 days to guard against the possibility that results are being driven by students with extreme intervals. Row 4 drops students who at some point during their repeated test administrations took the same form twice, which presumably would tend to inflate gains. Row 5 restricts the sample to the students who took the test exactly twice rather than at least twice because it is possible that students who took the test more than twice could be systematically different. Row 6 controls for the age of the student at the time of the initial administration to allow that students of different ages might have different average gains. Row 7 removes approximately 1,000 students from a single country that had identical values of interval of 184 days, because it seemed possible that this large block of students could be influencing the findings. Row 8 removes all students from the country where the average interval was extremely high (corresponding to the right-most mode in interval distribution in Figure 3). Row 9 removes all students from countries with fewer than 50 students. Row 10 removes students who were singletons in their testing group. The primary concern with these students is that because they did not appear to be on an institutional schedule, they may have been more likely to determine their own testing dates (hence motivating Model 2). Finally, Row 11 restricts the sample to students whose testing groups contained at least five students. The idea here was to isolate the sample to repeater students who clearly appeared to be part of a group testing effort initiated by their school or other program, because the smaller the testing group was, the more likely that the students could have had different educational experiences but just coincidentally were tested on the same dates under the same client. Again, the findings are robust.

The same is true for Model 2, as shown in Table A2. The table is analogous to Table A1, presenting the base model in the first row and then the same sensitivity analyses in following rows. The effects remain statistically significant at the .05 level, but overall the magnitude of the effects are somewhat reduced relative to Model 1. Interestingly, the final two rows indicate that the effect of interval after either removing singleton testing groups or restricting to larger testing groups is still evidently large. A possible explanation is that restricting to larger testing groups focuses the analysis on programs that might be better established or otherwise more organized at providing English instruction because they are evidently serving larger groups of students.

Table A1 Sensitivity Analyses Using Model 1 of the Effect of Interval on Total Gains

Row	Specification	*N*	Estimate	*SE*	*t*-stat.	*p*-value
1	Base	4,606	.045	.012	3.69	<.001
2	Include form pair dummy vars.	4,606	.042	.018	2.29	.022
3	Drop extreme intervals	4,591	.045	.012	3.65	<.001
4	Drop repeat form	4,425	.048	.013	3.62	<.001
5	Took test exactly twice	4,205	.050	.013	3.75	<.001
6	Control for student age	4,606	.044	.012	3.67	<.001
7	Drop interval of 184	3,599	.041	.012	3.34	.001
8	Drop country with large intervals	3,996	.046	.012	3.67	<.001
9	Drop countries with <50 students	4,459	.044	.012	3.56	<.001
10	Drop singleton testing groups	4,246	.056	.014	4.10	<.001
11	Restrict to testing groups with at least 5 students	3,770	.057	.017	3.42	.001

Table A2 Sensitivity Analyses Using Model 2 (Regression Adjustment for Baselines Scores) of the Effect of Interval on Total Gains

Row	Specification	*N*	Estimate	*SE*	*t*-stat.	*p*-value
1	Base	4,606	.030	.012	2.39	.017
2	Include form pair dummy vars.	4,606	.039	.018	2.16	.030
3	Drop extreme intervals	4,591	.030	.013	2.36	.018
4	Drop repeat form	4,425	.031	.013	2.34	.019
5	Took test exactly twice	4,205	.030	.013	2.26	.024
6	Control for student age	4,606	.029	.012	2.38	.017
7	Drop interval of 184	3,599	.028	.012	2.26	.024
8	Drop country with large intervals	3,996	.031	.013	2.43	.015
9	Drop countries with <50 students	4,459	.027	.013	2.18	.029
10	Drop singleton testing groups	4,246	.048	.014	3.42	.001
11	Restrict to testing groups with at least five students	3,770	.047	.017	2.81	.005

Suggested citation:

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). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12064

Action Editor: James Carlson

Reviewers: Guangming Ling and Yeonsuk Cho