Measuring the Power of Learning.®

Research Report
ETS RR–17-50

Evaluating the Stability of Test Score Means for the *TOEIC*® Speaking and Writing Tests

Yanxuan Qu

Yan Huo

Eric Chan

Matthew Shotts

December 2017

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 Stability of Test Score Means for the *TOEIC*® Speaking and Writing Tests

Yanxuan Qu, Yan Huo, Eric Chan, & Matthew Shotts

Educational Testing Service, Princeton, NJ

For educational tests, it is critical to maintain consistency of score scales and to understand the sources of variation in score means over time. This practice helps to ensure that interpretations about test takers' abilities are comparable from one administration (or one form) to another. This study examines the consistency of reported scores for the *TOEIC*® Speaking and Writing tests using statistical procedures. Specifically, the stability of the TOEIC Speaking score means from 431 forms administered in a 3-year period was evaluated using harmonic regression, and the stability of TOEIC Writing score means from 66 forms administered in a 3-year period was evaluated using analysis of variance. Results indicated that the fluctuations in the TOEIC Speaking or Writing score means mainly reflect changes in test takers' overall English speaking or writing ability levels instead of score inaccuracies. For both speaking and writing test scores, a large proportion of the variation in score means can be explained by seasonality (the rise or fall of score means associated with specific times of the year) and test takers' demographic information, which have been shown to be related to test-taker ability. As a result, this finding provides evidence for the consistency of the TOEIC Speaking and Writing score scales across forms.

Keywords Harmonic regression; ANOVA; seasonality; scale stability; quality control

doi:10.1002/ets2.12180

The *TOEIC*® Speaking and Writing tests are designed to measure a person's ability to communicate in spoken and written English, respectively, in the context of daily life and the global workplace. The TOEIC Speaking test is composed of 11 constructed-response questions and takes approximately 20 minutes to complete. The TOEIC Writing test is composed of eight constructed-response questions and takes approximately 1 hour to complete. Scores are reported on a scale of 0 to 200 with increments of 10 for both the speaking and the writing tests. Test takers can choose to take either the TOEIC Speaking test or the TOEIC Writing test or both. Both tests are administered on fixed dates at secure, Internet-based test centers. The TOEIC Speaking test is currently administered much more frequently than the TOEIC Writing test.

For tests with frequent administrations, it is of paramount importance that all score means be monitored over time. Evaluating the stability of test score means over time is an important quality control procedure to prevent errors in score reporting and to maintain test score validity by ensuring that the meaning of test scores is preserved. For a test to be valid, test scores must reflect the knowledge, skills, and abilities that the test is intended to measure. Kolen and Brennan (2014, p. 333) mentioned that one useful quality control procedure is checking the consistency of score statistics (e.g., score means and score variances) over time. When score statistics fluctuate, it is important to investigate the potential causes (Allalouf, 2007; von Davier, 2012). For example, the fluctuation of score means may be due to changes in test takers' demographic factors, seasonality (the rise or fall of score means associated with specific times of the year), the result of operational errors (e.g., errors in test score reporting), or test security breaches.

To better observe and monitor the pattern and trend of the many score means across different forms or administrations, researchers at Educational Testing Service (ETS) have used ANOVA and harmonic regression to check score mean fluctuations over time (Lee & von Davier, 2013; von Davier, 2012). For example, Haberman, Guo, Liu, and Dorans (2008) used the ANOVA method (Howell, 2002) to examine the stability of *SAT*® Math and Reading score means over a 9-year period. They found that the scales of SAT Math and Reading reporting scores were stable and the fluctuations in SAT score means were mainly due to seasonal effect. The ANOVA method was particularly appropriate given that the SAT test has a small number of forms a year with fixed schedules. Harmonic regression (Bloomfield, 2000) is appropriate when there are frequent numbers of administrations across the whole year so that the seasonality pattern in score means can be modeled by a smooth sinusoidal term in a time series manner (Lee & Haberman, 2013). Lee and Haberman (2013) used

Corresponding author: Y. Qu, E-mail: yqu@ets.org

the harmonic regression method to monitor score means across administrations for an international language test. They found that most of the fluctuations in the score means were explained by seasonal effect, yearly trend, and regional effect. Thus, the reporting scale for the language test was stable.

The purpose of this study was to evaluate the stability of the TOEIC Speaking and Writing test score means in an approximately 3-year period by using the harmonic regression method and the ANOVA method, respectively. Harmonic regression was chosen to monitor the stability of the TOEIC Speaking score means across forms due to the frequent administrations in Korea. Although the TOEIC Writing test was administered once a month in Korea, the number of forms in a year was sparse compared to those generated by the TOEIC Speaking test. Therefore, the ANOVA method was applied to check the stability of the TOEIC Writing score means across forms over time.

Data

The data for the TOEIC Speaking test were collected from Korean test takers who took only the TOEIC Speaking forms between February 1, 2014, and December 31, 2016. Background information was also available for each test taker (see the appendix for sample background questions). In total, 431 forms in 281 administrations were included in the analysis, with sample sizes ranging from 336 to 3,221 with an average sample size of 1,399. At the test administration level, sample sizes ranged from 336 to 11,022, with an average size of 2,135. The number of forms in each administration ranged from 1 to 5. Figure 1 shows the score means for all the 431 forms in a time series manner. The x-axis in Figure 1 is the number of days between each administration and January 1, 2014.

The data for the TOEIC Writing test contained writing scores and background information for Korean test takers who took forms with both speaking and writing sections between February 1, 2014, and December 31, 2016. We decided to use Korea-only data because (a) Korean test takers had the highest response rates to the background questionnaire and (b) Korean test takers regularly participated in the TOEIC Writing test (two forms each month on the same administration day) except after August 2016. In total, we had score data with background responses from 66 writing forms administered in Korea. Sample sizes per form ranged from 39 to 275, with an average sample size of 122.

Statistical Analyses

Harmonic regression is a linear regression model that contains sinusoidal terms. It can be used to check stability of score means because sinusoidal terms characterize seasonality in a time series fashion (Lee & Haberman, 2013). The harmonic regression models tried in this study are listed in Table 1.

In Table 1, where S_t is a mean score for Form t, Symbol d_t denotes the number of days elapsed since the beginning of 2014 and the time when Form t was administered. Symbol T_t is the total number of days in the year when Form t was administered. Year indicator $y_{1t} = 1$ indicates that Form t was administered in 2015, and $y_{2t} = 1$ indicates that Form t was administered in 2016. Score means in 2015 and 2016 were compared to score means in the baseline year, 2014.

Model 0 was a baseline model. Model 1 included the year effect terms. A significant year effect would indicate that the score means in year 2015 or 2016 were substantially higher (or lower) than in year 2014. Model 2 included sinusoidal terms for seasonal effect. Model 3 was a combined model with both year and seasonal effects. Model 4 is the complete model with year effect, seasonal effect and test takers' background effect. Four background variables were included in Model 4. After recoding of the original responses, f_{b3t} represented the fraction of test takers in each form who are not full-time employees, f_{b6t} represented the fraction of test takers in each form who have studied English for more than 10 years, f_{b8t} was the fraction of test takers in each form who used English more than 20% of the time in daily life, and f_{b10t} was the fraction of test takers in each form whose English did not always affect communication at work.

In our analyses, the year effect can be evaluated by comparing Model 1 to Model 0. The seasonality effect can be evaluated by comparing Model 2 to Model 0. The combined effect of year and seasonality can be evaluated by comparing Model 3 to Model 0. The combined effect of year, seasonality, and test takers' background can be evaluated by comparing Model 4 to Model 0. In our regression model, the seasonal terms and test takers' background variables are all indicators of test takers' performance on the test. As mentioned previously, seasonal factors are certain times of a year that are often related to business cycles within a year. Though related, seasonal factors and test takers' background factors are not necessarily identical.

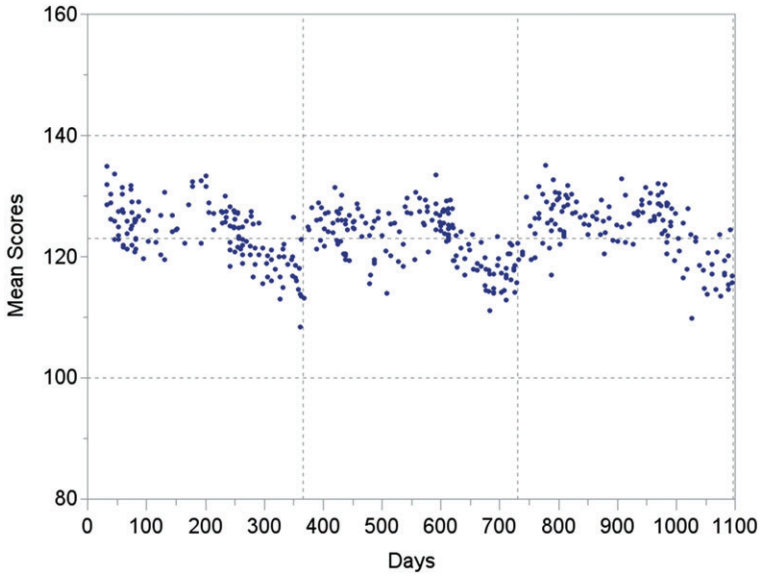

Figure 1 Mean TOEIC Speaking scores for 431 forms over time.

Table 1 Models for TOEIC Speaking Mean Scores

Model	Equation
Model 0	$S_t = \mu + e_t$
Model 1	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + e_t$
Model 2	$S_t = \mu + \beta_3 \cos(2\pi d_t/T_t) + \beta_4 \sin(2\pi d_t/T_t) + \beta_5 \sin(4\pi d_t/T_t) + \beta_6 \sin(6\pi d_t/T_t) + e_t$
Model 3	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + \beta_3 \cos(2\pi d_t/T_t) + \beta_4 \sin(2\pi d_t/T_t) + \beta_5 \sin(4\pi d_t/T_t) + \beta_6 \sin(6\pi d_t/T_t) + e_t$
Model 4	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + \beta_3 \cos(2\pi d_t/T_t) + \beta_4 \sin(2\pi d_t/T_t) + \beta_5 \sin(4\pi d_t/T_t) + \beta_6 \sin(6\pi d_t/T_t) + \beta_7 f_{b3t} + \beta_8 f_{b6t} + \beta_9 f_{b8t} + \beta_{10} f_{b10t} + e_t$

To determine which harmonic regression model was the best model and which terms could be added or dropped from the regression model, we followed Lee and Haberman's (2013) example and checked if the decrease in root mean square error (RMSE) was at least 5% after including the terms and if the increase in *R* square and adjusted *R* square was noticeable. Different from *R* square, adjusted *R* square evaluates model fit by taking into account the number of predictors in a model. Additionally, the residual plot was checked for model fit. To determine whether a regression coefficient was significantly different from zero, the *p* value of each regression coefficient in the final model was compared to 0.05 divided by the total number of predictors.

In the ANOVA analyses for the TOEIC Writing test, the dependent variable was the score mean for each writing form. The independent variables included month, year, and their interaction. In the final ANOVA model (Model 1), t represents form ($t = 1$ to 66), $\alpha_{m(t)}$ shows the seasonal effect, $\beta_{y(t)}$ shows the year effect, and $\delta_{m(t)\,y(t)}$ shows the interaction between month and year. We also included background variables in the ANOVA analyses. In Model 2, $\beta_{2b3(t)} + \beta_{3b6(t)} + \beta_{4b8(t)} + \beta_{5b10(t)}$ represents the effect from the four recoded background questions. As for the analyses of TOEIC Speaking scores, these four background variables were recoded into dummy variables.

Model 1: $M_t = \mu + \alpha_{m(t)} + \beta_{y(t)} + \delta_{m(t)\,y(t)} + e$.
Model 2: $M_t = \mu + \alpha_{\mathrm{m(t)}} + \beta_{y(t)} + \delta_{m(t)\,y(t)} + \beta_{2b3(t)} + \beta_{3b6(t)} + \beta_{4b8(t)} + \beta_{5b10(t)} + e$.

Results

Results for the *TOEIC*® Speaking Test

Table 2 shows that the *R* square value increased only slightly when the year indicator was added to the model (Model 1 vs. Model 0). However, adding the seasonal effect to the regression model increased *R* square significantly from 0.03 to

Table 2 Model Fitting Results: Number of Predictors, Root Mean Square Errors (RMSE), *R* Square, and Adjusted *R* Square

Model	Number of predictors	RMSE	R^2	Adjusted R^2
Model 0	0	4.8134	0	0
Model 1	2	4.7533	0.0294	0.0248
Model 2	4	3.4803	0.4821	0.4772
Model 3	6	3.4206	0.5020	0.4950
Model 4	10	3.2230	0.5621	0.5517

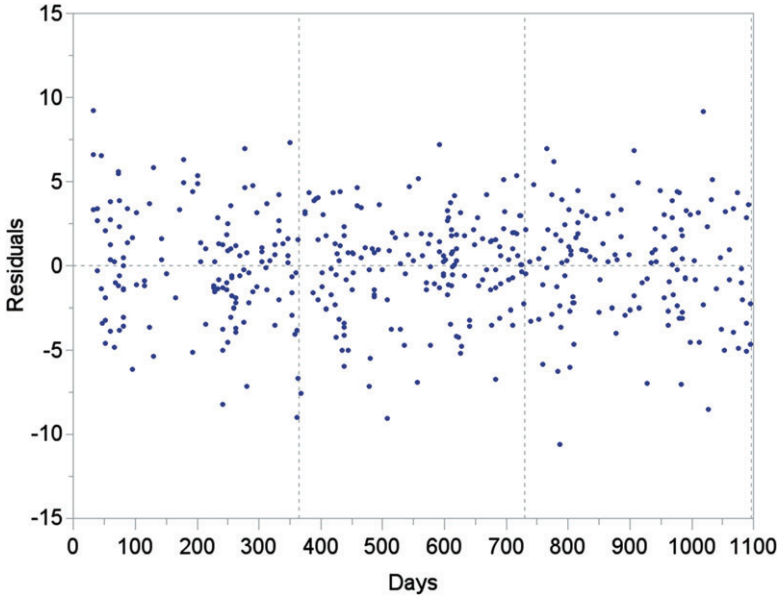

Figure 2 Residuals for 431 TOEIC Speaking test forms over time.

0.48. Adding test takers' background information to the regression model also increased the amount of explained variation and decreased the amount of unexplained error noticeably. From Model 3 to Model 4, *R* square increased from 0.50 to 0.56, by almost 12%, and RMSE decreased from 3.4206 to 3.2230, by 5.8%. Model 4 was chosen as the final model because no other indicators were found that could decrease RMSE by more than 5%. The fit of Model 4 was checked by a residual plot (Figure 2). Residuals are the difference between observed score means and predicted score means. All the residuals for the 431 forms appeared to be randomly and evenly distributed in Figure 2, indicating appropriate model fit.

Table 3 shows the parameter estimates for the final model (Model 4). Since we conducted significance tests for 10 predictors simultaneously in the regression model, the *p* values of each predictor were compared to $0.05/10 = 0.005$. A *p* value less than 0.005 indicates that the predictor is statistically significant. Therefore, the parameter estimates for the two year indicators, β_1 and β_2, were not significant, indicating very small score mean variations across 3 years. At least two seasonal parameters (β_3 andβ_5) had a *p* value less than 0.005, which means the score means followed a strong periodical pattern over time. This periodical pattern can be seen clearly in Figure 1. Three background variables also had significant parameter estimates. These background variables were the fraction of test takers in each form who had studied English for more than 10 years, who used English more than 20% of the time in daily life, and whose English did not always affect communication at work.

Figure 3 shows both observed (denoted by dots) and predicted (denoted by plus signs) mean scores for all 431 forms by the number of days elapsed between their administration date and January 1, 2014. A periodic pattern is clearly seen. In each year, the mean scores were relatively higher around the end of the first quarter and the third quarter but lower in the fourth quarter. This seasonal pattern is quite similar across the 3 years. There were multiple predicted values at an administration day in Figure 3 because there were multiple forms in one administration day.

Table 3 Estimated Parameters in Model 4 (The Final Model)

Model	Parameter	Estimate	*SE*	*T* statistic	*p* value
y_{1t}	β_1	−0.7304	0.3919	−1.8600	0.0631
y_{2t}	β_2	0.4475	0.4317	1.0400	0.3005
$\cos(2\pi d_t/T_t)$	β_3	−1.5334	0.3802	−4.0300	<.0001
$\sin(2\pi d_t/T_t)$	β_4	0.7782	0.2967	2.6200	0.0090
$\sin(4\pi d_t/T_t)$	β_5	2.1295	0.4622	4.6100	<.0001
$\sin(6\pi d_t/T_t)$	β_6	0.0353	0.2421	0.1500	0.8840
f_{b3t}	β_7	6.1553	2.9104	2.1100	0.0350
f_{b6t}	β_8	30.1261	8.1743	3.6900	0.0003
f_{b8t}	β_9	31.1538	9.1552	3.4000	0.0007
f_{b10t}	β_{10}	52.8702	12.3195	4.2900	<.0001

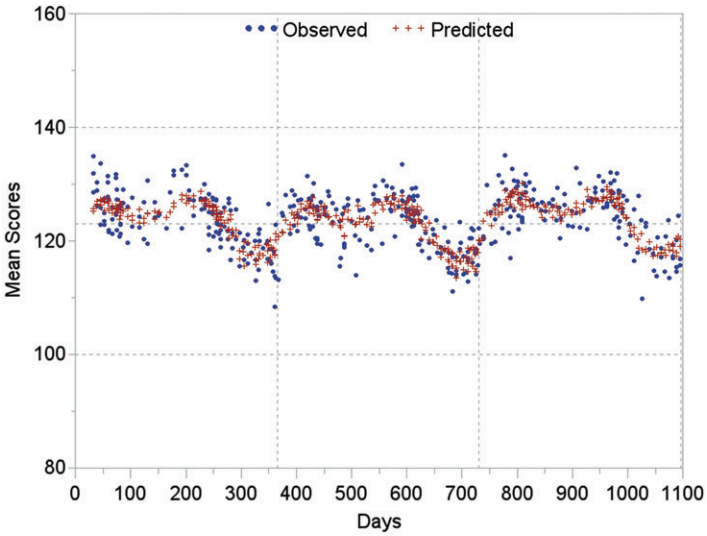

Figure 3 Observed and predicted TOEIC Speaking score means for 431 forms over time.

Results for the *TOEIC*® Writing Test

Tables 4 and 5 summarize the numbers of TOEIC Writing forms and the means and standard deviations of score means by month and by year. For example, Table 4 shows that there were six writing forms administered in July across 3 years. The average of these reported score means was 146.98 and the standard deviation was 3.83. Typically, two writing forms were administered each month, however, only one writing form was administered in September, October, November, and December during 2016. As a result, the total number of forms across 3 years was five instead of six in these 4 months in Table 4, and the total number of forms in 2016 was 20 instead of 24 in Table 5. Table 5 also shows that 22 instead of 24 writing forms were administered in 2014 in Korea because our data did not have forms administered in January 2014.

Unlike the results for speaking, adding the four background variables did not reduce RMSE or increase *R* square substantively. In fact, RMSE increased only from 3.52 to 3.7, and *R* square increased from 0.81 to 0.82. None of the parameter estimates for the four background variables was statistically significant. Therefore, the final model did not include any background variables. Table 6 shows the ANOVA results for the final model (Model 1).

Table 6 indicates that month was the major variable accounting for the score mean variations. It explained 41% of the total mean score variance. Figure 4 indicates that the average score means (connected by solid lines in Figure 4, with circles representing the score means for individual forms) tended to be higher in the first and third quarters than in the second and fourth quarters. This pattern bears some resemblance to the periodic pattern observed in the speaking results.

Different from the ANOVA results for speaking, the year effect was significant for writing, and so was the interaction effect. Table 5 shows that the average score means in 2014 and 2015 were similar to each other, whereas the average score mean of 2016 was higher than the other 2 years, especially in February, May, and December (as seen in Figure 4). However,

Table 4 Summary Statistics of TOEIC Writing Score Means by Month of Administrations

Month	*N*	Mean	*SD*	Min	Max
January	4	151.61	3.07	148.26	155.52
February	6	153.51	5.81	144.36	161.03
March	6	149.22	5.86	139.87	157.37
April	6	146.50	3.49	142.24	152.20
May	6	142.22	4.83	134.64	147.92
June	6	143.12	1.76	140.32	145.00
July	6	146.98	3.83	141.15	151.55
August	6	148.27	4.15	142.35	153.82
September	5	147.93	2.57	143.63	149.90
October	5	142.11	4.08	135.26	145.35
November	5	143.09	5.62	135.68	147.53
December	5	140.99	4.98	136.50	149.15
Overall	66	146.30	5.54	134.64	161.03

Table 5 Summary Statistics of TOEIC Writing Score Means by Year of Administrations

Month	*N*	Mean	*SD*	Min	Max
2014	22	144.49	4.98	135.26	153.25
2015	24	145.57	5.85	134.64	158.25
2016	20	149.18	4.80	139.87	161.03
Overall	66	146.30	5.54	134.64	161.03

Table 6 ANOVA Results for TOEIC Writing Scaled Scores (Based on Individual Form Level: $N = 66$, Total $R^2 = 0.81$)

Component	*df*	Sum of squares	Mean square	*F*	*p*	R^2
Month	11	808.97	73.54	5.95	<.0001	0.41
Year	2	173.90	86.95	7.04	0.003	0.09
Interaction	21	492.10	23.43	1.9	0.05	0.25
Residual	31	383.03	12.36			0.19

the score means for September 2016 through December 2016 were only based on one form. More data cumulated over a longer time period would be needed to better understand if there is indeed a year effect or an interaction effect between year and month of the administrations for writing.

Concluding Remarks

The results based on harmonic regression for speaking showed significant seasonal effect and demographic effect. In all 3 years, the TOEIC Speaking score means appeared to be higher around March and August and lower in the other months. Given the large number of forms, the large number of administrations for the TOEIC Speaking test each year, and the sample size per form, the regression model explained a reasonably high proportion (56%) of total mean score variation across forms. A large portion of the observed fluctuation in test score means was explained by the fact that test takers differ systematically in their ability and demographic characteristics according to the time of the year they choose to take the test (seasonal effects). It can be argued, therefore, that the scale of the TOEIC Speaking test is appropriately stable, after accounting for seasonal and demographic differences in test takers' overall speaking ability.

The results for writing also showed a significant seasonal effect. Within each year, the score means appeared to fluctuate in a pattern similar to the one detected for the TOEIC Speaking test. Overall, the ANOVA model explained 81% of the total variation in writing score means. The scale of the TOEIC Writing test is also stable.

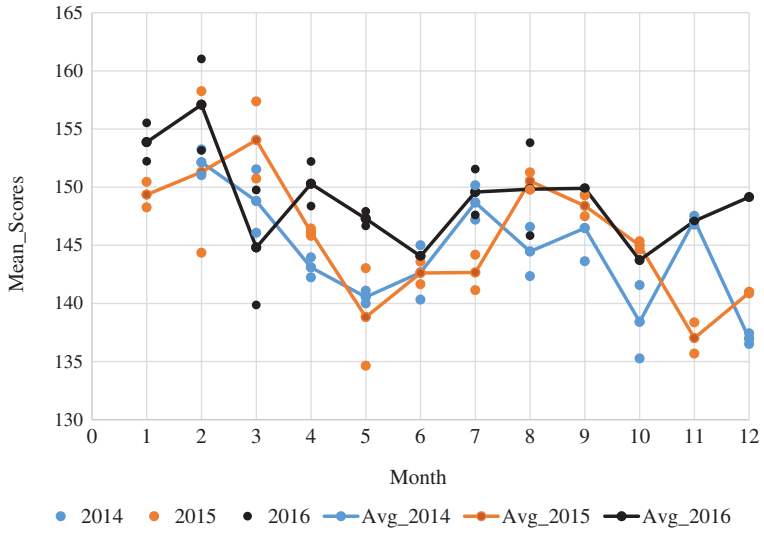

Figure 4 Writing score means by year and by month.

References

Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26*(1), 36–46.

Bloomfield, P. (2000). *Fourier analysis of time series: An introduction* (2nd ed.). New York, NY: Wiley.

Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning test score conversions* (Research Report No. RR-08-67). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02153.x

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*(4), 815–829.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*(3), 557–575. https://doi.org/10.1007/s11336-013-9317-5

von Davier, A. A. (2012). *The use of quality control and data mining techniques for monitoring scaled scores: An overview* (Research Report No. RR-12-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02302.x

Appendix

Example Background Questions

Which of the following best describes your current status?

01. I am employed full-time (including self-employed).
02. I am employed part-time and/or study part-time.
03. I am not employed.
04. I am a full-time student.

How many years have you spent studying English?

01. Less than or equal to 4 years
02. More than 4 years but less than or equal to 6 years
03. More than 6 years but less than or equal to 10 years
04. More than 10 years

How much time must you use English in your daily life?

01. None at all
02. 1 to 10%
03. 11 to 20%

04. 21 to 50%
05. 51 to 100%

How often has difficulty with English affected your ability to communicate?

01. Almost never
02. Seldom
03. Sometimes
04. Frequently
05. Almost always

Suggested Citation:

Qu, Y., Huo, Y., Chan, E., & Shotts, M. (2017). *Evaluating the stability of test score means for the* TOEIC® *Speaking and Writing tests* (Research Report No. RR-17-50). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12180

Action Editor: Donald Powers

Reviewers: Ru Lu and Jiahe Qian