JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12266 *Journal of Educational Measurement*

Studying Score Stability with a Harmonic Regression Family: A Comparison of Three Approaches to Adjustment of Examinee-Specific Demographic Data

Yi-Hsuan Lee
Educational Testing Service
Shelby J. Haberman
Consultant

For assessments that use different forms in different administrations, equating methods are applied to ensure comparability of scores over time. Ideally, a score scale is well maintained throughout the life of a testing program. In reality, instability of a score scale can result from a variety of causes, some are expected while others may be unforeseen. The situation is more challenging for assessments that assemble many different forms and deliver frequent administrations per year. Harmonic regression, a seasonal-adjustment method, has been found useful in achieving the goal of differentiating between possible known sources of variability and unknown sources so as to study score stability for such assessments. As an extension, this paper presents a family of three approaches that incorporate examinees' demographic data into harmonic regression in different ways. A generic evaluation method based on jackknifing is developed to compare the approaches within the family. The three approaches are compared using real data from an international language assessment. Results suggest that all approaches perform similarly and are effective in meeting the goal. The paper also discusses the properties and limitations of the three approaches, along with inferences about score (in)stability based on the harmonic regression results.

Introduction

It is common practice for assessments to use different test forms in different administrations. Whenever possible, proper designs are made to permit equating of test scores to adjust form difficulty in different administrations and to ensure comparability of test scores over time (Kolen & Brennan, 2014). Ideally a score scale is well maintained throughout the life of a testing program. In reality, many reasons may cause inconsistencies in a score scale. Haberman and Dorans (2011) identified six sources of variability in score distributions that may contribute to score-scale inconsistency: test-construction practices, subpopulation shifts and changes in populations, sampling errors, accumulation of random equating error, the role of the anchor, and model misfit. When equating works satisfactorily for an assessment, the mean examinee scores of individual administrations still may change within each year—by month, by season, and so on—due to cyclical seasonal shifts in subpopulations that can be fairly consistent over years (e.g., Haberman, Guo, Liu, & Dorans,

Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

2008; Lee & Haberman, 2013; Liu & Yoo, 2019; Qu, Huo, Chan, & Shotts, 2017). Such within-year seasonality is considered in this paper. For assessments that assemble many different test forms and deliver frequent administrations in a year, it can be especially challenging to maintain a stable score scale because, in addition to potential shifts in (sub)populations, more variations may occur that lead to scale inconsistency. For example, with frequent administrations, the sample size per administration is more difficult to guarantee and the sample may not be representative of the target population of the assessment; such sampling error has impact on the accuracy of equating (Haberman, 2010). Also, when a large number of test forms need to be equated successively, statistical errors in equating methods may cumulate quickly and the accumulated errors can produce scale drift (Guo, 2010; Haberman, 2010; Livingston, 2004). Allalouf, Gutentag, and Baumer (2017) emphasized the importance of quality control in testing and discussed possible errors that may occur in different stages of the testing process. Thus, there is a strong need to monitor test scores over time and study possible changes in the score scale on a regular basis.

In the literature, a number of studies have been conducted to address this need. The first type of approaches relies on special equating designs to be planned in advance to assess scale drift directly (e.g., Petersen, Cook, & Stocking, 1983; Puhan, 2008). For example, Puhan (2008) considered two equating designs, one with parallel chains and the other with a single long chain; for each equating design, scale stability was assessed by comparing the equating functions produced from different chains. The second type of approaches utilizes current and historical data of an assessment in a statistical model to identify possible sources of administration-to-administration variability in test scores that may be easily explained, and then assess the unexplained variability in the scores (e.g., Haberman et al., 2008; Lee & Haberman, 2013; Lee & von Davier, 2013; Lee, Liu, & von Davier, 2014; Liu & Yoo, 2019; Qu et al., 2017; Wei & Qu, 2014). Thorough evaluation of the identified sources of variability and the degree and patterns of the unexplained variability can then suggest score (in)stability.

Among the second type of approaches mentioned above, some analyzed examinee-level data (e.g., Lee et al., 2014; Wei & Qu, 2014) and others analyzed administration-level data (e.g., Haberman et al., 2008; Lee & Haberman, 2013; Lee & von Davier, 2013; Liu & Yoo, 2019; Qu et al., 2017). The former approaches studied the effects of examinee-specific demographic data on individual scores, including a linear mixed-effects model applied to 15 administrations structured by a special equating design (Lee et al., 2014) or a multilevel analysis applied to 254 administrations of a test in 4 years (Wei & Qu, 2014). Both studies examined many demographic variables[1] but did not consider any types of seasonality in test scores. Among the latter approaches, Haberman et al. (2008) examined data from 54 administrations in 9 years of SAT®, using a two-way analysis of variance (ANOVA) model to examine monthly effects (i.e., seasonality), yearly trends, and their interactions in the mean scores and the raw-to-scale conversions. As the model involves fixed effects for month of administration, this approach is more useful for assessments that are administered a few times a year and the administration schedule is regular (i.e., for SAT, 6 fixed months every year but different dates across years). Lee and Haberman (2013) proposed a harmonic regression approach to analyze mean examinee scores of 211 administrations in 3 years. Harmonic regression employs sinusoidal functions

to characterize seasonal patterns in data, and additional administration-level predictors may be included for adjustment of yearly trends and changes in the regional distribution of examinees in mean scores across administrations. As the sinusoidal functions are continuous, this approach is more suitable for assessments that deliver very frequent administrations, even with a relatively short history of stable operation. For example, if the seasonal patterns in the mean scores of an assessment repeat over a 1-year period, then data from 1 year of administrations under the same testing conditions and operational procedures should suffice for an initial harmonic regression analysis, before data from more seasonal cycles under the same operation can be accumulated for updated analyses. This harmonic regression approach has been successfully applied to such tests as the TOEIC® Speaking test (Qu et al., 2017) and the GRE® General Test (Liu & Yoo, 2019).

The harmonic regression approach in Lee and Haberman (2013) only involves administration-level data: It is a one-stage approach where the response variable is mean examinee scores of individual administrations and the predictors are terms relevant to the time of administration and administration-level demographic data (e.g., fraction of examinees taking the test in a certain region in an administration). When the examinees who take a particular administration of an assessment are not a random sample of all examinees who take the assessment (i.e., the target population of the assessment), the demographic distribution of the examinees may vary substantially across administrations. This phenomenon has been found in many references mentioned above with different assessments. To further enhance the applicability and utility of harmonic regression in different operational settings, it may be beneficial to extend the existing harmonic regression approach to make additional adjustment using examinee-specific demographic data. To date, no research has explored this extension.

This paper presents a family of approaches in which examinees' demographic data are incorporated into harmonic regression in three different ways. It also develops a generic evaluation method to compare the different approaches within the family. As will be shown, the introduced harmonic regression family contributes to score monitoring by supplying an efficient tool for separating possible known sources of variability in scores from the unknown sources—the objective of this study is to examine what type of harmonic regression approaches can achieve this goal more effectively, given the same set of examinee demographic data. The approach that yields smaller unexplained variability in the administration-level scores indicates a better method to tease out possible known sources of variability in the score distributions. The modeling results should then be reviewed thoroughly to infer scale (in)consistency for an assessment and actions to be taken if necessary.

The introduced family includes the Lee and Haberman (2013) approach, which is treated as the benchmark in our comparison because it has been successfully applied to different assessments in the literature (Liu & Yoo, 2019; Qu et al., 2017). We also demonstrate how one may define the predictors related to the regional distribution of examinees somewhat differently to improve the performance of the original Lee and Haberman (2013) approach. The second and third approaches use different methods to adjust for changes in the examinee-specific demographic data before assessing the stability of mean scores at the administration level. One approach applies harmonic

regression directly to study the relationship between examinees' scores and their demographic data and seasonality. Its idea is similar to those in Lee et al. (2014) and Wei and Qu (2014), but a simpler model is employed herein with seasonality considered. The other approach involves adjustment of the examinees' demographic distribution by the minimum discriminant information adjustment (MDIA; Haberman, 1984). MDIA is a general procedure for weighting samples. In educational measurement, MDIA has been employed to produce pseudo-equivalent groups for linking test scores across administrations without proper anchor items (Haberman, 2015; Lee, Haberman, & Dorans, 2019). In our study, MDIA is applied to mitigate variations in examinees' demographic distributions among individual administrations, so that the (in)consistency of mean scores can be assessed based on more homogeneous examinee samples across administrations.

The remainder of this paper is structured as follows. The "Methods" section defines the three harmonic regression approaches statistically, and develops the generic evaluation method to compare models using the jackknife (Wolter, 2007, chapter 4). The techniques are applied to real data from an international language assessment in the "Application" section. The paper concludes with a discussion about the findings, the properties and limitations of the three approaches, the inferences about score (in)stability based on harmonic regression results, and possible scenarios and the recommended actions if score instability is discovered. An Appendix is provided in the end to demonstrate how to express the three harmonic regression approaches in matrix notation to facilitate the development of the jackknife results in the "Methods" section.

Methods

This section describes the notation used for the harmonic regression family, introduces the three specific approaches considered in this paper, and discusses the evaluation method for model comparisons. As mentioned earlier, the different approaches share the same goal: They are intended to remove possible sources of variation of the scores that may be easily explained, and then assess the extent of the unexplained variability in the scores. However, they differ in how the possible sources of variation are modeled in regression using the same set of examinee demographic information. It is noteworthy that equating, as well as the assessment of scale drift, is usually accomplished at the administration level, so the main source of variation is administration. Thus, the three approaches are compared based on the unexplained variability in the scores at the administration level. For those comparisons, the sample size that matters is the number of administrations. The sampling variation due to the finite number of observed examinees within each administration is relatively small. In the end, the approach that yields smaller unexplained variability in the administration-level scores, or the root mean squared error (RMSE) of prediction, indicates a better approach to teasing out possible known sources of variability in the score distributions.

The first approach, referred to as *Average Score*, focuses on the average examinee scores in individual administrations and is a one-stage approach. It is adopted from Lee and Haberman (2013) and treated as the benchmark in the comparison.

The second approach, termed *Average Residual*, has two stages. It begins by merging the examinee data from different administrations together and applying harmonic regression to study variations in the examinee scores due to differences in the examinees' demographic variables. Then the residuals from the examinee-level harmonic regression are averaged by administration for further evaluation and comparison. The third approach, named *Weighted Average*, is also a two-stage analysis. It first applies MDIA (Haberman, 1984) to weight the administration samples one by one so that the sample characteristics become more similar to the target population of the assessment. Next, for individual administrations, the weighted averages of the examinee scores based on the MDIA weights are treated as a time series, and then analyzed by harmonic regression to further remove possible sources of variation that are not accounted for in MDIA.

In the application of harmonic regression, sinusoidal functions are used to characterize seasonal patterns in the response variables, which typically repeat every year in assessments (Lee & Haberman, 2013). To relate administration t to sinusoidal functions, variables associated with the time of administration are needed. Define variable d_t as the number of days elapsed since the beginning of the year at the time of administration t. The year-length variable L_t is equal to 365 for an ordinary year, and is equal to 366 for a leap year. For instance, an administration given on January 10, 2014, has $d_t = 10$ and $L_t = 365$. The kth harmonic component is expressed as $a_k \cos(2\pi k d_t/L_t) + b_k \sin(2\pi k d_t/L_t)$, where a_k and b_k are two unknown coefficients to be estimated for each harmonic component. For all three approaches, the models include the first K harmonic components to account for seasonality,

$$\sum_{k=1}^{K}[a_k \cos(2\pi k d_t/L_t) + b_k \sin(2\pi k d_t/L_t)]. \tag{1}$$

This paper considers common seasonal patterns across regions and other demographic variables. That is because the within-year cyclical seasonal shifts in subpopulations are strongly influenced by academic and professional calendars that lead to specific deadlines for submission of test results that must be met and are not that variable for different examinees. One could add interactions for seasonality by region or other demographic variables in the adjustment, but the impact on the modeling results tends to be small in practice.

Approach 1: Average Score

This approach is based on the harmonic regression model introduced in Lee and Haberman (2013), which analyzes administration-level data. The response variable of this model is the sample mean of individual scores (or mean score) for each administration. Consider T administrations of an assessment given to $N = \sum_{t=1}^{T} N_t$ total examinees, with N_t being the number of examinees in administration t, $1 \le t \le T$. The data from the T administrations are combined to form the target population of the assessment. For examinee n in administration t, let Y_{tn} be the test score. Denote the mean score of administration t, $1 \le t \le T$, as $\bar{Y}_t = N_t^{-1} \sum_{n=1}^{N_t} Y_{tn}$. Predictors include variables that describe seasonality in the data and variables that characterize changes in the regional distribution of examinees, although it may be sensible to

incorporate additional predictors into harmonic regression to capture different patterns in the mean scores and to describe changes in the demographic distribution of the examinees due to other sources. Let $F_{t,r} \geq 0$ be the fraction of examinees who took administration t in region r, $1 \leq r \leq R$, where R is the total number of regions defined for the target population of the assessment and $\sum_{r=1}^{R} F_{t,r} = 1$. For instance, if 10% of the examinees in Administration 1 were in Region 2, then $t = 1$, $r = 2$, and $F_{1,2} = 0.1$. In this paper, the fractions $F_{t,r}$, $1 \leq r \leq R$, generally describe the regional distribution of the examinees in administration t. Among the R regions, assume that Region R has the most examinees overall in the T administrations, although the order of regions is arbitrary. In Lee and Haberman (2013), the regions were defined as the top five testing countries and others, so that $R = 6$. Other grouping variables related to the regional distribution of the examinees may be worth considering. As discussed in the "Application" section, this study considers a new grouping variable that utilizes information about both the region and the language background of the examinees, and $R = 14$ in this new definition.

In general, a harmonic regression model with the first K harmonic components to account for seasonality and with the region fractions to account for changes in the regional distribution of examinees is given by

$$\text{Model 1:} \quad \bar{Y}_t = \mu_1 + \sum_{k=1}^{K} [a_k \cos(2\pi k d_t / L_t) + b_k \sin(2\pi k d_t / L_t)] + \sum_{r=1}^{R-1} c_r F_{t,r} + e_{1,t}, \tag{2}$$

where $e_{1,t}$ are independent random variables with common mean zero and variance σ_1^2,[2] and μ_1, a_k and b_k ($1 \leq k \leq K$), and c_r ($1 \leq r \leq R-1$) are unknown constants to be estimated. This study uses the largest region, Region R, as the reference for regional adjustment to ensure more stable regression results than using a smaller region as the reference. Thus, Model 1 in Equation 2 leaves out $F_{t,R}$ for $1 \leq t \leq T$, and only includes $R-1$ region fractions; their coefficients c_r, $1 \leq r \leq R-1$, may be interpreted as follows: Relative to the average score of the reference region, increasing 1% in the fraction $F_{t,r}$ of region r in administration t is expected to increase the average score $\bar{Y}_t$ of administration t by $0.01c_r$. Additional predictors can be added to Model 1 in Equation 2 in the same way as the harmonic components or the region fractions. The base model for Model 1 is

$$\bar{Y}_t = \mu_1 + e_{2,t}, \tag{3}$$

which has no predictor.

The overall predictive value of Model 1 and its base model is primarily assessed through RMSE, R^2, and adjusted R^2. Adjusted R^2 is considered because it combines information about prediction error with the number of parameters. Many residual diagnostics may be considered to further evaluate the model fit, as elaborated in Draper and Smith (1998) and Lee and Haberman (2013). In this study, we focus on one particular residual diagnostic, which assesses the assumption that the random errors in the harmonic regression models are independent across administrations.

The Durbin-Watson statistic is a standard test for serial correlation (Draper & Smith, 1998, pp. 181–192). A large serial correlation has several implications and merits special attention: It may indicate that more harmonic components are needed for seasonal adjustment, or may relate to a large linear trend that should also go into the model. It may be introduced if linking of test scores uses consecutive administrations (e.g., Li, Li, & von Davier, 2011). In addition, the serial correlations may arise due to errors in linking procedures that are consistent (Haberman & Dorans, 2011). Time-series methods that handle serial correlations, such as the class of autoregressive-integrated moving average (ARIMA) models (Brockwell & Davis, 1991, chapter 9), may be considered if the Durbin-Watson test result is significant but adding more harmonic components or additional linear trends does not help.

Average Score (Approach 1) only requires aggregated data. That can be an advantage for researchers without access to individual-level information. Also, testing programs may have summary statistics of test scores and of demographic variables related to proficiency readily available for ease of monitoring, so that application of this approach can have basically no cost in data preparation.

Approach 2: Average Residual

As mentioned earlier, this approach has two stages. At the first stage, examinees from all T administrations are merged and analyzed at once, using a harmonic regression model to study the relationship between examinee scores Y_{tn}, $1 \le n \le N_t$ and $1 \le t \le T$, and such predictors as examinee demographic variables $\boldsymbol{X}_{tn}$ and harmonic components. For each examinee n in administration t, $\boldsymbol{X}_{tn}$ is a d-dimensional vector of demographic variables. The study concerns categorical demographic variables (e.g., region, gender, age group, etc.), and they are coded as dummy variables. The examinee-level harmonic regression model (Model 2e) with the first K harmonic components may be expressed as

$$\text{Model 2e:} \quad Y_{tn} = \mu_2 + \boldsymbol{\alpha}' \boldsymbol{X}_{tn} + \sum_{k=1}^{K} [a_k \cos(2\pi k d_t / L_t) + b_k \sin(2\pi k d_t / L_t)] + e_{3,tn}, \tag{4}$$

where $\boldsymbol{\alpha}$ is a d-dimensional coefficient vector for the predictors $\boldsymbol{X}_{tn}$. An ANOVA approach may be considered for Model 2e, which takes the individual predictors as different sources of variance and investigates how they contribute to the overall model R^2. To serve this purpose, the overall R^2 for Model 2e is evaluated, together with the semipartial ω^2 (SAS Institute Inc., 2018, p. 4034) estimated for each predictor in the model that partitions the overall R^2. The model yields a predicted score $\hat{Y}_{tn}$ for examinee n in administration t, and the ordinary residual $r_{tn} = Y_{tn} - \hat{Y}_{tn}$ is computed for $1 \le n \le N_t$ and $1 \le t \le T$. The examinee residuals are then aggregated by administration; the sample mean of the residuals (or mean residual), $\bar{r}_t = N_t^{-1} \sum_{n=1}^{N_t} r_{tn}$, is computed for administration t, $1 \le t \le T$. The overall sample mean of the examinee residuals r_{tn} across all examinees and administrations is equal to 0, but the sample mean of the mean residuals $\bar{r}_t$ across administrations is different from zero. The reason is that sample sizes N_t, $1 \le t \le T$, vary by administration. As changes

in weights $(1/N_t)$ in computing the sample mean of $\bar{r}_t$ across t have remarkably low impact in typical cases, the difference from 0 is small.

At the second stage, the mean residuals $\bar{r}_t$, $1 \leq t \leq T$, are treated as the response variable in the following administration-level model:

$$\text{Model 2a:} \quad \bar{r}_t = \mu_3 + e_{4,t}. \tag{5}$$

The RMSE of this model is equivalent to the standard deviation (*SD*) of $\bar{r}_t$ across t. Similarly to *Average Score* (Approach 1), the Durbin-Watson test is conducted for Model 2a in Equation 5 to detect if there is a large serial correlation among the administration-level scores that remains after adjusting for seasonality and changes in the demographic variables.

It is worth noting that the adjustment of seasonality and of the demographic distribution of examinees is accomplished at the individual level in this approach, so the administration-level model in Equation 5 has no predictor. As the sample size at the individual level is typically large, it is often feasible to incorporate as many predictors as are available. This is in contrast to the administration-level harmonic regression model considered in *Average Score* (Approach 1), where the smaller number of data points for individual administrations implies that one has to worry more about variable selection and likely fall back on rather limited predictors.

Approach 3: Weighted Average

Like *Average Residual* (Approach 2), this approach also has two stages. At the first stage, the demographic distribution of examinees is adjusted using MDIA (Haberman, 1984), which obtains a weighted sample of examinees for each administration with sample characteristics more similar to the target population of the assessment. The consistency of the mean scores of the weighted examinee samples is then analyzed at the second stage with administration-level harmonic regression models.

In MDIA, weights w_{tn} are obtained for each administration t and examinee n so that, for administration t, (a) the sum of the weights in an administration is $\sum_{n=1}^{N_t} w_{tn} = 1$, and (b) the weighted average $\sum_{n=1}^{N_t} w_{tn} \boldsymbol{X}_{tn}$ of the demographic variables for the administration is equal to the overall average $\bar{\boldsymbol{X}} = N^{-1} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \boldsymbol{X}_{tn}$ of the demographic variables across all administrations (i.e., this is the target population defined in the study). Among all weights w_{tn} satisfying the two constraints, for each administration t, the MDIA weights $\hat{w}_{tn}$, $1 \leq n \leq N_t$, are chosen to minimize the Kullback-Leibler discriminant information $\sum_{n=1}^{N_t} w_{tn} \log(N_t w_{tn})$ for comparison of the weights w_{tn} to uniform weights with constant value $1/N_t$. Suppose that the demographic variables $\boldsymbol{X}_{tn}$, $1 \leq n \leq N_t$, have a positive-definite sample covariance matrix. For unique positive $\hat{c}_t$ and d-dimensional vector $\hat{\boldsymbol{\beta}}_t$, the MDIA weights have an exponential form $\hat{w}_{tn} = \hat{c}_t \exp(\hat{\boldsymbol{\beta}}_t' \boldsymbol{X}_{tn})$ for $1 \leq n \leq N_t$. By use of the Newton-Raphson algorithm, the $\hat{\boldsymbol{\beta}}_t$ can be found by solving the equation from constraint (b)

$$\bar{\boldsymbol{X}} - \sum_{n=1}^{N_t} \hat{w}_{tn}^{*} \boldsymbol{X}_{tn} = 0 \tag{6}$$

with $\hat{w}^*_{tn} = \exp(\hat{\boldsymbol{\beta}}_t' \boldsymbol{X}_{tn})$, and then $\hat{c}_t = 1/\sum_{n=1}^{N_t} \hat{w}^*_{tn}$ due to constraint (a). The estimation procedure is detailed in Haberman (1984). Software is available via a license for noncommercial use (Haberman, 2014).

Once the weighted sample is available for each administration, at stage 2, the weighted mean scores $\bar{Y}_{w,t} = \sum_{n=1}^{N_t} \hat{w}_{tn} Y_{tn}$ are computed and modeled with a harmonic regression with the first K harmonic components,

$$\text{Model 3:} \quad \bar{Y}_{w,t} = \mu_4 + \sum_{k=1}^{K} [a_k \cos(2\pi k d_t / L_t) + b_k \sin(2\pi k d_t / L_t)] + e_{5,t}. \quad (7)$$

Changes in the regional distribution of examinees are already adjusted at stage 1 of the approach along with other demographic variables, so their effects are not considered in Model 3 in Equation 7. The base model for Model 3 is

$$\bar{Y}_{w,t} = \mu_4 + e_{6,t}, \quad (8)$$

which has no predictor. As in *Average Score* (Approach 1), the overall evaluation of Model 3 and its base model is based on RMSE, R^2, and adjusted R^2. The Durbin-Watson test for serial correlation is also considered for these two administration-level models.

Compared to *Average Score* (Approach 1), the two-stage approaches—that is, *Average Residual* (Approach 2) and *Weighted Average* (Approach 3)—have the same advantage that the adjustment of the examinees' demographic distribution is accomplished with examinee-level data, so it is possible to consider a large number of demographic variables in the adjustment. It is worth noting that MDIA attempts to match the demographic distribution of examinees in an administration to the demographic distribution of the target population through weighting. It may not work well if there are demographic subgroups with very small proportions (or even zero proportion) in an administration relative to their proportions in the target population. Thus, instead of using those demographic subgroups directly in MDIA, they should be combined with other subgroups that are similar in test performance for the adjustment.

Comparison of Approaches in the Harmonic Regression Family

Comparison of the three approaches is based on RMSE of the final administration-level models. In our study, the models to be compared are Model 1 in Equation 2, Model 2a in Equation 5, and Model 3 in Equation 7. As noted earlier, this evaluation is done with respect to mean scores/residuals because the assessment of scale drift is usually accomplished at the administration level. Smaller RMSE indicates less variability in mean scores across administrations after accounting for seasonality and changes in the regional distribution, as well as other demographic variables for *Average Residual* (Approach 2) and *Weighted Average* (Approach 3); that implies better results for the adjustment of the examinees' demographic distribution. As differences in the RMSE for different approaches may be due to sampling errors, the delete-one jackknife method (Wolter, 2007, chapter 4) is applied to facilitate the comparison.

This subsection develops the jackknife results using the matrix approach to linear regression and externally studentized residuals (e.g., Draper & Smith, 1998), as well

as standard formulas for jackknife estimates (Wolter, 2007, chapter 4). In general, consider the following regression model in matrix notation,

$$\mathbf{V} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}. \tag{9}$$

The T-dimensional response vector $\mathbf{V}$ has elements V_t, $1 \leq t \leq T$ for administrations; its prediction is denoted as $\hat{V}_t$. The $\mathbf{Z}$ is a $T \times p$ matrix with rank $p < T - 1$, which contains a T-dimensional column of ones for the intercept and $p - 1$ columns of dimension T for predictors. The $\boldsymbol{\gamma}$ is a p-dimensional coefficient vector. The $\mathbf{e}$ is a T-dimensional error vector with elements e_t, $1 \leq t \leq T$ for administrations, and the e_t are independent random variables with mean zero and variance σ^2. The general matrix expression in Equation 9 covers all of the administration-level models to be compared. Readers who are interested in the detailed model specifications are referred to the Appendix.

To develop the jackknife results, suppose that $\hat{e}_t = V_t - \hat{V}_t$ is the ordinary residual for administration t. Let SSE be the residual sum of squares with $T - p$ degrees of freedom, and denote the residual mean square as MSE. Then $\text{MSE} = \text{SSE}/(T - p)$, and $\text{RMSE} = \sqrt{\text{MSE}}$ (Draper & Smith, 1998, chapter 4). To apply the delete-one jackknife method to RMSE, the technique used to develop the externally studentized residuals (Draper & Smith, 1998, pp. 207–208) for the model in Equation 9 may be employed. Let $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ be the $T \times T$ hat matrix, and let $H_{tt'}$ be the row t and column t' element of $\mathbf{H}$. The observation V_t contributes 1 degree of freedom of residual sum of squares $m_t = \hat{e}_t^2/(1 - H_{tt})$ to the overall SSE. Then removal of observation V_t reduces SSE to $\text{SSE}_{(t)} = \text{SSE} - m_t$ with $T - p - 1$ degrees of freedom (Draper & Smith, 1998, pp. 207–208). Accordingly, the MSE changes to

$$\text{MSE}_{(t)} = \frac{1}{T - p - 1}\text{SSE}_{(t)} = \text{MSE} + \frac{1}{T - p - 1}(\text{MSE} - m_t). \tag{10}$$

Thus, leaving out observation V_t reduces $\text{RMSE} = \sqrt{\text{MSE}}$ to $\text{RMSE}_{(t)} = \sqrt{\text{MSE}_{(t)}}$.

With the $\text{MSE}_{(t)}$ and $\text{RMSE}_{(t)}$, standard formulas for the jackknife method may be applied to produce the bias-corrected RMSE and the SE of RMSE (Wolter, 2007, pp. 152–153). It is noteworthy that MSE is an unbiased estimate of σ^2 if the regression model holds (Draper & Smith, 1998, p. 140). In this case, RMSE is a biased estimate of σ unless RMSE is a constant: $E(\text{MSE}) = E(\text{RMSE}^2) = [E(\text{RMSE})]^2 + \text{Var}(\text{RMSE})$, so $\sigma \leq E(\text{RMSE})$ and the equality holds only when $\text{Var}(\text{RMSE}) = 0$. In addition, MSE is not unbiased if the predictors are random and the regression model does not necessarily hold exactly. Thus, RMSE is generally not unbiased, and the bias may be reduced using the jackknife. The tth pseudo-value for RMSE is $\text{RMSE}_t = T\text{RMSE} - (T - 1)\text{RMSE}_{(t)}$, and the jackknife estimate of RMSE is

$$\text{RMSE}_{JK} = \frac{1}{T}\sum_{t=1}^{T}\text{RMSE}_t \tag{11}$$

(Wolter, 2007, p. 152). The RMSE_{JK} is a bias-corrected RMSE, and is referred to as the jackknife RMSE in the "Application" section.

Table 1
Summary Statistics of Individual Examinee Scores Per Test Section

Test Section	*N*	Mean	*SD*
Reading	361,746	19.785	6.778
Listening	361,746	20.738	6.576
Speaking	361,746	21.326	4.467
Writing	361,746	20.075	4.890

To produce the SE of RMSE, let s_m^2 be the sample variance of the m_t. The jackknife variance estimate for MSE is

$$\text{Var(MSE)} = \frac{T-1}{T}\sum_{t=1}^{T}(\text{MSE}_{(t)} - \overline{\text{MSE}}_{(\cdot)}), \tag{12}$$

where $\overline{\text{MSE}}_{(\cdot)} = T^{-1}\sum_{t=1}^{T}\text{MSE}_{(t)}$ is the sample mean of $\text{MSE}_{(t)}$ (Wolter, 2007, p. 153). Given Equations 10 and 12, it can be found that

$$\text{Var(MSE)} = \frac{(T-1)^2}{T(T-p-1)^2}s_m^2. \tag{13}$$

As $\text{RMSE} = \sqrt{\text{MSE}}$, use of the delta method (Casella & Berger, 2002, p. 243) yields

$$\text{Var(RMSE)} = \frac{1}{4\text{MSE}}\text{Var(MSE)} = \frac{(T-1)^2}{(4\text{MSE})T(T-p-1)^2}s_m^2. \tag{14}$$

The Var(RMSE) is the jackknife variance estimate for RMSE, and the jackknife SE is $\sqrt{\text{Var(RMSE)}}$.

Application

Data

In the study, all administrations from an international language assessment of English proficiency were used such that the administration date was between August 2013 and August 2014, and the test was given primarily in western countries to at least 3,000 examinees. The assessment delivered almost weekly administrations using different test forms. It had four test sections (Reading, Listening, Speaking, and Writing) and they were scaled separately. In all, the study involved $T = 36$ administrations and $N = 361{,}746$ examinees. The data from the 36 administrations were merged to construct the target population of the assessment we studied. For each examinee, the following information was available: the unrounded scale scores of the four test sections, the country in which they took the test (i.e., testing country), and demographic information such as age, gender, native country, and native language. Table 1 shows the summary statistics (mean and *SD*) of individual scores by test section.

Demographic information regarding gender, native country, and native language was gathered from background questionnaires and involved a small number of

missing responses. The missing data were treated as a separate category for each demographic variable. Examinee's date of birth was required for test registration, so age was available for everyone. The examinees were classified into subgroups based on demographic variables. Four age groups were defined: below 18 (10.85%), between 18 and 22 (37.27%), between 23 and 30 (37.99%), or above 30 (13.89%). There were three gender groups: male (49.51%), female (49.30%), and missing responses about gender (1.18%).

As examinee performance on the language assessment has been found to correlate with their language background (Lee & Haberman, 2016), we considered two grouping variables related to the regional distribution of the examinees. One objective of our study was to compare which of the two grouping variables was able to account for more variability in the examinee scores. In Lee and Haberman (2013), the examinees were classified into six regions solely based on the testing countries: they were the top five testing countries and others available in a data set. This is the first grouping variable we used, still termed *region* herein following the original work. Applying the same grouping variable to our data set led to such regions as Brazil, France, Germany, Turkey, the United States, and others. The actual regions are different from those in Lee and Haberman (2013) due to the use of a different data set.

The second grouping variable utilized information about testing countries and native countries, with consideration of the target population of the assessment we studied. For this assessment, native country and native language revealed similar information about the examinees' language background, so we only used one of them in the analyses. This new rule, termed *region-language combination* henceforth, led to a finer classification of the target population of the assessment. It is expected to classify the target population into more homogeneous subgroups because many of the examinees took the test in foreign or English-speaking countries (e.g., the United States) with different native languages, and grouping them together solely based on testing countries (like the regions in the first grouping variable) may create less homogeneous subgroups with respect to language background. On the other hand, region and language of the examinees we studied are closely related, so they cannot be cross-classified for everyone. It implies that a regression model with region and native country as separate predictors is likely to have multicollinearity issues. Thus, we chose to define a new grouping variable that accounts for both region and language of the examinees and use as one set of predictors. In total, the following 14 region-language combinations were defined: Africa (i.e., examinees whose native country is in Africa and took the test in Africa), Africa abroad (i.e., examinees whose native country is in Africa but took the test outside Africa), Americas, Americas abroad, China abroad, Europe, Europe abroad, India abroad, Japan abroad, Korea abroad, Middle East, Middle East abroad, Asia abroad (i.e., examinees from Asian countries other than China, India, Japan, and Korea but took the test outside their native country), and English-speaking countries (i.e., examinees who are not in the above regions and took the test in English-speaking countries or Pacific islands). It is worth noting that Americas and Americas abroad only included examinees whose native language is not English.

Among the demographic variables available in the data set, regional distribution of the examinees was the sole demographic variable used in all three approaches. Gender and age group were used in *Average Residual* (Approach 2) and *Weighted Average* (Approach 3). For each demographic variable, the largest subgroup was treated as the reference group to ensure more stable regression results. They were the following: age—ages between 23 and 30, gender—male, region—others, and region-language combination—Europe.

Results

The three approaches were applied to the data described above. The performance of each approach was examined first with intermediate results discussed. For the final evaluation, the RMSE of the final administration-level models and the associated jackknife estimates were then considered for the comparison of different approaches. Table 2 provides an overview of the analyses and key models involved in each of the three approaches and the final evaluation, as well as the corresponding result tables.

Recall that for *Average Score* (Approach 1), the mean score of each administration was computed, and then analyzed as a time series in Model 1 and its base model. The time-series plot in Figure 1 shows the mean scores of the four test sections for the 36 administrations by the administration day relative to August 1, 2013. Table 3 presents summary statistics of the mean scores across administrations. Preliminary analysis indicated that Model 1 in Equation 2 with the first two harmonic components ($K = 2$) tended to perform better than did the model with $K = 1$ or $K = 3$ for all test sections, so that the model with $K = 2$ was chosen to further compare which of the two grouping variables related to the regional distribution of the examinees (i.e., region versus region-language combination) could explain more variations in the mean scores. Table 4 shows the results of overall model evaluation and of the Durbin-Watson test for three models for a given test section: (a) the base model for Model 1 in Equation 3 (no predictor), (b) Model 1 in Equation 2 with two harmonic components (four predictors) and the fractions of regions (five predictors), and (c) Model 1 in Equation 2 with two harmonic components (four predictors) and the fractions of region-language combinations (13 predictors). The model with the region-language combination explained considerably more variation in the mean scores of all test sections in terms of RMSE, R^2, and adjusted R^2, and hence outperformed the existing region definition proposed in Lee and Haberman (2013). The adjusted R^2 were comparable and above .785 for all test sections with predictors based on the region-language combination. Thus, the 14 region-language combinations were used in the analysis for *Average Residual* (Approach 2) and *Weighted Average* (Approach 3) for adjustment of the regional distribution of the examinees. For all test sections, the Durbin-Watson test was not statistically significant (i.e., above .05) for any models with harmonic components (Model 1). It indicates that, once the two harmonic components were included in the model to account for seasonality, there was no significant serial correlation that remained and needed additional treatment. Thus, other models that may address serial correlation in average scores (e.g., harmonic regression with a linear trend or ARIMA models) were not necessary for our data. Model 1 was deemed the final model for *Average Score* (Approach 1) for further comparisons.

Table 2
An Overview of the Analyses, Key Models, and Result Tables for the Three Approaches and the Final Evaluation

Approach	Stage 1 Examinee Level	Stage 2 Administration Level	Result Table		Note
			Stage 1	Stage 2	
Average Score		Model 1, Region		Table 4	Benchmark
Average Score		Model 1, Region-language		Table 4	New
Average Residual	ANOVA, Model 2e	Model 2a	Table 5		New
Weighted Average	MDIA	Model 3	Table 6	Table 7	New
Final evaluation		Jackknife, four models above		Table 8	

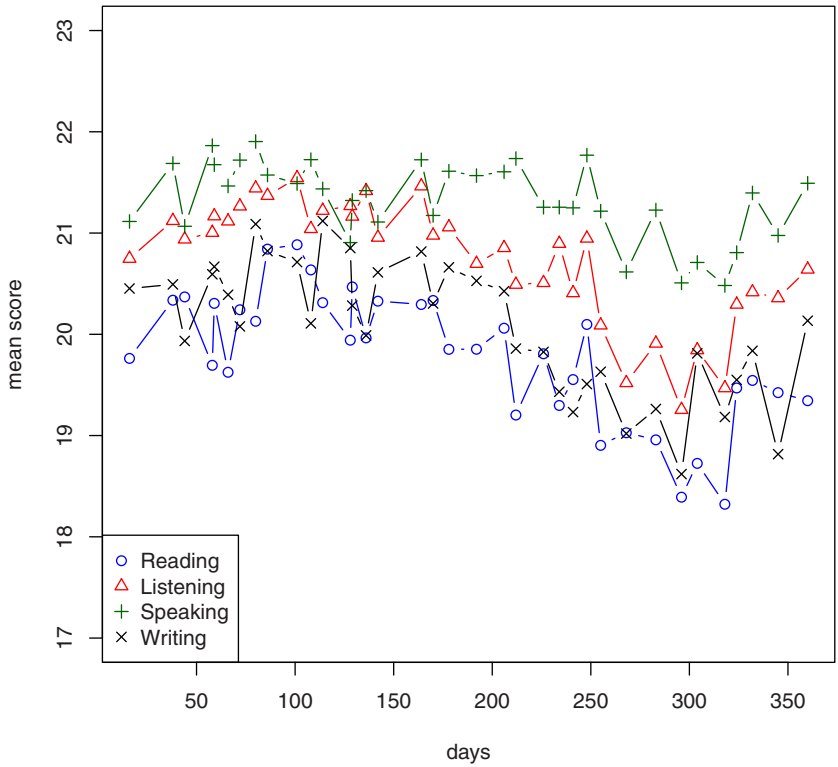

Figure 1. Time-series plot of mean scores for the four test sections. (Color figure can be viewed at wileyonlinelibrary.com)

Table 3
Summary Statistics of Mean Scores Per Test Section

Test Section	*N*	Mean	*SD*
Reading	36	19.787	.644
Listening	36	20.747	.594
Speaking	36	21.329	.383
Writing	36	20.074	.650

As noted in the "Methods" section, *Average Residual* (Approach 2) begins with ANOVA with the examinee scores and then studies the mean residuals from the ANOVA model across administrations. At stage 1, Model 2e in Equation 4 was applied to the data with the first two harmonic components ($K = 2$) and the three demographic variables available—gender, age group, and region-language combination. The analysis involved 22 predictors for each test section excluding the reference group for each demographic variable (i.e., two for gender, three for age group, 13 for region-language combination, and four for the two harmonic components). As mentioned earlier, ANOVA was performed for the individual examinee scores per

Table 4
Harmonic Regression Results for Average Score (Approach 1)

Test Section	Model	Number of Predictors	RMSE	R^2	Adjusted R^2	Durbin-Watson Test Stat	Durbin-Watson p-Value
Reading	Base model for Model 1	0	.644	.000	.000	.612	.000
	Model 1, Region	9	.352	.778	.701	1.937	.301
	Model 1, Region-language	17	.281	.902	.809	2.238	.718
Listening	Base model for Model 1	0	.594	.000	.000	.425	.000
	Model 1, Region	9	.288	.826	.765	2.396	.687
	Model 1, Region-language	17	.258	.903	.811	2.142	.508
Speaking	Base model for Model 1	0	.383	.000	.000	1.220	.014
	Model 1, Region	9	.218	.760	.677	2.307	.902
	Model 1, Region-language	17	.157	.914	.833	2.705	.247
Writing	Base model for Model 1	0	.650	.000	.000	.745	.000
	Model 1, Region	9	.363	.768	.688	2.309	.898
	Model 1, Region-language	17	.302	.889	.785	2.647	.338

test section to assess the impact of the different predictors (as sources of variability) present in the data. See Table 5 for the results. In the ANOVA, effects for seasonality were decomposed into four sources in Model 2e in Equation 4: sin and cos for the first harmonic component, $\sin(2\pi d_t/L_t)$ and $\cos(2\pi d_t/L_t)$; and sin2 and cos2 for the second harmonic component, $\sin(4\pi d_t/L_t)$ and $\cos(4\pi d_t/L_t)$. Generally, the overall R^2 for the model was comparable (about .1) for all test sections. This level of overall R^2 is not surprising when predicting individual examinees' test scores using their demographic information; for similar findings from different assessments, see, for example, Lee et al. (2014) and Wei and Qu (2014). Region-language combination was the most important source of variation in the scores. Gender had the least impact on all test section scores except for Speaking. The contribution of the harmonic components in terms of R^2 ranged from .004 to .013 across test sections.

At stage 2 of *Average Residual* (Approach 2), for each test section, the mean residual $\bar{r}_t$ from Model 2e in Equation 4 was computed by administration for $1 \leq t \leq 36$; then Model 2a in Equation 5 was fitted to the mean residuals for the 36 administrations. The RMSE of the administration-level models (Model 2a) were equal to .354 for Reading, .266 for Listening, .268 for Speaking, and .349 for Writing. As a comparison, consider the results in Table 4 for *Average Score* (Approach 1). The RMSE for Model 2a was similar to the corresponding RMSE for Model 1 with six regions. More important, for all test sections, there was a considerable decrease in RMSE from the base model for Model 1 to Model 2a. Thus, adjusting for seasonality and for the composition of gender, age group, and region-language combination in the examinee-level data (stage 1) indeed reduced the variability in scores at the aggregated administration level, even though the examinee-level model at stage 1 had a low overall R^2. The Durbin-Watson test for serial correlation was conducted for Model 2a. The p-value of the Durbin-Watson test was above .2 for all test sections, indicating insignificant serial correlation in the mean residuals. Thus, the detailed Durbin-Watson test results are not further tabulated. Model 2a was the final model for *Average Residual* (Approach 2) for further comparisons.

For *Weighted Average* (Approach 3), MDIA was employed for each administration to obtain a weight for each examinee, so that the weighted sample of the examinees had comparable demographic distributions in terms of gender, age group, and region-language combination to the target examinee population. The weights were then applied to produce the weighted sample mean $\bar{Y}_{w,t}$ of the scale scores for each test section of an administration. Table 6 presents summary statistics of the weighted mean scores based on MDIA. Compared to the summary statistics of the original mean scores (Table 3), the weighted mean scores were slightly greater on average and more variable for all test sections. The base model for Model 3 in Equation 8 (no predictor) and the harmonic regression Model 3 in Equation 7 with the first two harmonic components $K = 2$ (four predictors) were applied to the weighted mean scores $\bar{Y}_{w,t}$. See Table 7 for the results. Among the four test sections, Model 3 was most effective for Listening and least effective for Speaking in terms of (adjusted) R^2. Again, the Durbin-Watson tests were not statistically significant for all models with harmonic components (Model 3), so there was no clear serial correlation that remained in the residuals from Model 3. For further comparisons, the final model for *Weighted Average* (Approach 3) was Model 3.

Table 5
ANOVA of Individual Examinee Scores by Test Section for Average Residual (Approach 2, Model 2e)

Test Section	Source	DF	SS	MS	F-Value	R^2	Overall R^2
Reading	sin	1	56,397.2	56,397.2	1,391.4	.003	.118
	cos	1	56,067.5	56,067.5	1,383.3	.003	
	sin2	1	2,117.0	2,117.0	52.2	.000	
	cos2	1	1,509.3	1,509.3	37.2	.000	
	Gender	2	8,620.1	4,310.1	106.3	.001	
	AgeGrp	3	143,830.2	47,943.4	1,182.8	.009	
	Region-language	13	1,689,762.4	129,981.7	3,206.8	.102	
	Residual	361,723	14,661,603.0	40.5			
Listening	sin	1	49,860.8	49,860.8	1,281.8	.003	.101
	cos	1	52,175.2	52,175.2	1,341.3	.003	
	sin2	1	5,741.7	5,741.7	147.6	.000	
	cos2	1	916.6	916.6	23.6	.000	
	Gender	2	3,721.3	1,860.7	47.8	.000	
	AgeGrp	3	8,8756.9	2,9585.6	760.6	.006	
	Region-language	13	1,371,230.9	105,479.3	2,711.5	.088	
	Residual	361,723	14,071,141.6	38.9			

(*Continued*)

Table 5
Continued

Test Section	Source	DF	SS	MS	F-Value	R^2	Overall R^2
Speaking	sin	1	5,580.0	5,580.0	312.8	.001	.106
	cos	1	10,661.2	10,661.2	597.6	.001	
	sin2	1	6,395.5	6,395.5	358.5	.001	
	cos2	1	5,268.5	5,268.5	295.3	.001	
	Gender	2	79,123.2	39,561.6	2,217.6	.011	
	AgeGrp	3	36,625.1	12,208.4	684.3	.005	
	Region-language	13	622,407.1	47,877.5	2,683.7	.086	
	Residual	361,723	6,453,193.4	17.8			
Writing	sin	1	53,680.4	53,680.4	2,499.4	.006	.102
	cos	1	49,497.3	49,497.3	2,304.6	.006	
	sin2	1	5,959.3	5,959.3	277.5	.001	
	cos2*	1	55.4	55.4	2.6	.000	
	Gender	2	29,842.9	14,921.5	694.7	.003	
	AgeGrp	3	41,487.4	13,829.1	643.9	.005	
	Region-language	13	702,138.0	54,010.6	2,514.7	.081	
	Residual	361,723	7,768,942.8	21.5			

Note. DF is degrees of freedom, SS is sum of squares, MS is mean square, and R^2 is the contribution to the overall R^2 of the source based on the estimated semipartial ω^2. Sources "sin" and "cos" compose the first harmonic component. Sources "sin2" and "cos2" compose the second harmonic component. AgeGrp is age group. Region-language is region-language combination. (*) For this predictor (i.e., cos2 for Writing), the p-value for the F-test was equal to .108. It is the only case in this table with a p-value greater than 10^{-6}.

Table 6
Summary Statistics of Weighted Mean Scores Per Test Section Based on MDIA

Test Section	*N*	Mean	*SD*
Reading	36	19.816	.677
Listening	36	20.769	.640
Speaking	36	21.350	.425
Writing	36	20.089	.652

With the final administration-level models for the three approaches, now the focus shifts to their comparison. As mentioned earlier, the different approaches were compared based on the RMSE from the final administration-level models, and the delete-one jackknife method was applied to account for sampling errors in the RMSE. Table 8 presents the jackknife results for different test sections and approaches. Similarly to Table 4, this table also shows two final models for *Average Score* (Approach 1), one for Model 1 with region and the other for Model 1 with region-language combination. The results for Model 1 with region are included as the benchmark to represent the application of the original Lee and Haberman (2013) approach to our data. For *Average Score*, it is already mentioned that using the finer region-language combinations rather than the more coarse regions in Model 1 explained more variability in the mean scores for all test sections. When comparing the final models for the three different approaches, it is clear that using region-language combination in *Average Score* (Approach 1) led to the smallest RMSE for all test sections and outperformed the two new approaches. *Average Residual* (Approach 2) and *Weighted Average* (Approach 3) yielded similar RMSE for the administration-level models for every test section, with *Average Residual* (Approach 2) performing slightly better. Both new approaches were competitive with *Average Score* (Approach 1) with region (i.e., the original Lee & Haberman, 2013, approach). The bias-corrected jackknife estimates of RMSE showed the same pattern. When taking the SE into consideration, the differences in (jackknife) RMSE across approaches and models were not substantial for all test sections except for Speaking. In summary, all three approaches performed similarly on Reading, Listening, and Writing. On Speaking, *Average Score* with the finer region-language combinations was most effective in adjusting for seasonality and for changes in the demographic distribution of examinees with the current data.

Discussion

The paper presents a family of harmonic regression approaches that can be used to investigate score stability for very frequently administered assessments, even with a relatively short history of stable operation. It also develops a generic evaluation method to compare different approaches in the family. This harmonic regression family includes the original Lee and Haberman (2013) approach and two new approaches that utilize examinees' demographic data in different ways. The three approaches are applied to and compared using data from four test sections of an international language assessment. With the current data, results suggest that *Average Score*

Table 7
Harmonic Regression Results for Weighted Average (Approach 3)

Test Section	Model	Number of Predictors	RMSE	R^2	Adjusted R^2	Durbin-Watson Test Stat	Durbin-Watson *p*-Value
Reading	Base model for Model 3	0	.677	.000	.000	.556	.000
	Model 3	4	.366	.741	.708	2.148	.752
Listening	Base model for Model 3	0	.640	.000	.000	.370	.000
	Model 3	4	.274	.837	.816	2.167	.796
Speaking	Base model for Model 3	0	.425	.000	.000	.995	.001
	Model 3	4	.280	.615	.566	2.357	.740
Writing	Base model for Model 3	0	.652	.000	.000	.695	.000
	Model 3	4	.369	.717	.680	2.392	.658

Table 8
Jackknife Estimate of RMSE and Jackknife SE for RMSE for the Three Approaches

Test Section	Approach	Note	Model	Number of Predictors	RMSE	Jackknife RMSE	Jackknife SE
Reading	Average Score	Region	Model 1	9	.352	.348	.071
	Average Score	Region-language	Model 1	17	.281	.301	.052
	Average Residual		Model 2a	0	.354	.357	.039
	Weighted Average		Model 3	4	.366	.366	.041
Listening	Average Score	Region	Model 1	9	.288	.281	.047
	Average Score	Region-language	Model 1	17	.258	.266	.052
	Average Residual		Model 2a	0	.266	.268	.027
	Weighted Average		Model 3	4	.274	.276	.031
Speaking	Average Score	Region	Model 1	9	.218	.216	.037
	Average Score	Region-language	Model 1	17	.157	.153	.031
	Average Residual		Model 2a	0	.268	.270	.028
	Weighted Average		Model 3	4	.280	.283	.035
Writing	Average Score	Region	Model 1	9	.363	.364	.045
	Average Score	Region-language	Model 1	17	.302	.311	.082
	Average Residual		Model 2a	0	.349	.352	.039
	Weighted Average		Model 3	4	.369	.374	.050

(Approach 1) with a new grouping variable that considers both testing countries and native countries and leads to a finer classification of the examinee population performed better on Speaking than the two new approaches (i.e., *Average Residual* and *Weighted Average*) and the original Lee and Haberman (2013) approach (*Average Score* with region). The results for the other three test sections are comparable for the different approaches. Overall, all three harmonic regression approaches are found effective in differentiating between possible known sources of variability and the unknown sources.

The current findings may relate to characteristics of the assessment and the data studied. In our study, a very small number of demographic variables are available and the missing rate is low. Region-language combination is the most important source of variation in the scores, and its effects are accounted for by all three approaches. The two new approaches, *Average Residual* and *Weighted Average*, take the additional demographic variables (age and gender) into consideration, but these variables tend to have less variation across administrations and are less predictive of test performance. Thus, it is not surprising that further adjustment using these demographic variables does not improve the results with our data.

A promising finding is that all three harmonic regression approaches, in conjunction with different grouping variables, can be useful in monitoring score scales. These approaches do not assume that everything is accounted for by the included predictors. It essentially provides a measure of variability not explained by the regression, rather than variability that might be explained. Obviously, it is best if the available information does explain most of the variability, which would then be indicative of score stability.

The harmonic regression family allows researchers and practitioners to choose and compare among the different approaches according to the characteristics of their assessment data, with considerations about the different properties and limitations of the three approaches: *Average Score* (Approach 1) has been found to perform well in this study and several existing studies (Lee & Haberman, 2013; Liu & Yoo, 2019; Qu et al., 2017). It is easy to include different predictors to characterize a variety of patterns and to describe changes in the examinees' demographic distribution in the data. For example, beyond regional effects, Qu et al. (2017) considered predictors relating to the examinees' employment status and experience of English learning and usage in daily life for the TOEIC Speaking test, while Liu and Yoo (2019) included such background characteristics as language status, educational status, and graduate major objective in the study for the GRE General Test. As it analyzes administration-level data, the method usually works better with a larger number of administrations that span more seasonal cycles under the same operation—for instance, about 71 administrations per year for 3 years in Lee and Haberman (2013), about 144 administrations per year for 3 years in Qu et al. (2017), and about 24 administrations per year for 5 years in Liu and Yoo (2019). This feature somewhat constrains the number of predictors that can be included in this harmonic regression model. *Average Residual* (Approach 2) is the most flexible and potentially most widely applicable method among the three approaches examined. Sample size is typically not an issue because examinee-level data are used for the adjustment. Predictors in Model 2e, including administration-specific variables and examinee-specific variables, can be

easily defined and incorporated for different assessments no matter how many demographic variables are available. As in *Average Residual* (Approach 2), sample size is not an issue for *Weighted Average* (Approach 3). MDIA can be accomplished with any examinee-specific variables in which missing categories are combined with other categories that are similar in test performance for the adjustment. However, it is trickier to employ administration-specific variables in MDIA, which is why seasonality is adjusted at stage 2 of the approach in the study. It is worth noting that all regions and region-language combinations existed in every administration in the data set. For assessments that deliver administrations in nonoverlapping regions over time, MDIA (and hence *Weighted Average*) does not apply unless the analysis is conducted by region. Due to these properties and limitations, different harmonic regression approaches are likely to be adequate for different assessments. For instance, *Average Residual* (Approach 2) and *Weighted Average* (Approach 3) are likely to be more successful if there are demographic data that are more predictive of the examinees' test scores. On the other hand, demographic information is usually self-reported and not mandatory, so it may not be accurate and the missing rate may be high for some assessments. When it is difficult to collect or use demographic information or when this information is of uncertain quality, one may wind up with relatively simple harmonic regression models that only involve mandatory demographic information so that the different adjustment approaches may yield similar results. To better understand the benefit of these harmonic regression approaches, further investigation is needed with data from different assessments.

Harmonic regression is considered a quality-control procedure. For monitoring of a score scale, it is desirable to accumulate data and update an existing harmonic regression analysis on a routine basis. With more data over time, the harmonic regression is expected to capture the systematic patterns in the test scores more accurately, and the resulting RMSE may better represent the unexplained variability in the score distribution. If the updated harmonic regression analyses look substantially different from time to time, then unexpected changes may have occurred in the assessment data that merit investigation. In every application, it is crucial to identify a model that best differentiates between possible known sources of variability in the score distributions and the unknown sources. The best models may differ across applications; examining the differences would help understand how the score distributions may have changed over time. The present study focuses on the comparison of different approaches in the harmonic regression family. The detailed residual diagnostics, model interpretations, and the use of residuals to detect unusual administrations described in Lee and Haberman (2013) are applicable to all models in the family so they are not repeated herein. After fitting the models, work is needed to interpret the findings with respect to scale (in)consistency. The identified sources of variability are subject to evaluation, in terms of whether such changes across administrations can be interpreted in light of the test design and delivery. For instance, (sub)population shifts are a common source of variability that may introduce seasonality in the score distributions, which is easy to verify due to the cyclical patterns within each year (Haberman & Dorans, 2011). This should also be a relatively stable source of variability in scores when the harmonic regression analysis is updated with more data from the same assessment. Changes in the examinees' regional distribution may be revealed

in exploratory analysis. Yearly increases, or long-term trends, in test scores may be less clear to interpret because they can result from true improvement in performance or scale/item drift. To tease out one effect from the other, it may be worthwhile to conduct a special equating design to examine scale drift specifically (e.g., Petersen et al., 1983; Puhan, 2008); methods for detecting item drift (Bock, Muraki, & Pfeiffenberger, 1988; Donoghue & Isham, 1998; Guo, Robin, & Dorans, 2017; Zhang & Li, 2016) may be considered as well. If there are different assessments with similar target populations, then true improvement in performance is likely observed in more than one of these assessments.

The extent of the unexplained variability provides a general picture about score (in)stability. Excess residual variance from a properly chosen model may indicate large variability above equating sampling error; for example, due to sampling of anchor items (Haberman, Lee, & Qian, 2009; Michaelides & Haertel, 2014), accumulation of random equating error after a chain of equatings (Guo, 2010; Haberman, 2010; Livingston, 2004), or other sources of variability (see, e.g., Haberman & Dorans, 2011). Patterns in the residuals are also worth examining to detect additional time structures or drift that should also be accounted for in the models (e.g., Lee et al., 2011; Lee & von Davier, 2013), which may be informed by a large serial correlation detected by the Durbin-Watson test. That in turn is potentially indicative of a security breach or a scaling/equating problem that would need to be remedied with additional investigation and statistical analyses. Additional quality-control procedures (Allalouf et al., 2017) may be used to detect different types of errors in different stages of the testing process.

Once score instability is discovered, it is important to figure out what is happening. Here are some possible scenarios and the recommended actions:

- In very severe security breaches, it is important to treat the security problem and to redo equating for a substantial period of time based on cleaned data.
- In case of systematic equating errors over long periods, correction of errors, redoing equating, and scale changes are generally involved.
- In some cases, there really is a major change in the proficiency of examinees. In this case, test redesign and rescaling may be needed.
- In some cases, the equating is incorrect inherently. Changes in equating methodology may be needed to obtain more effective linkage.
- In some cases, the equating is not incorrect. It is just not adequate due to poorly performing anchors or too few anchors. The proper solution involves test redesign to permit better anchors.
- In some cases, equating has been deliberately distorted due to client pressure and lack of institutional integrity. Solutions generally demand honesty, redoing equating, and rescaling.
- Changes in method of test administration may lead to substantial changes in general performance and in anchor behavior. This problem is important in transitions from paper-based testing to computer-based testing and in transitions from computers to tablets. It may be necessary to gradually redo equating and rescale.

- Once in a while, operations for a specific administration involve substantial errors but nothing else is involved. The results for the administration need to be fixed to the extent possible.

Recall that the "Introduction" section mentions two types of statistical models that make use of current and historical data of an assessment under stable operation to differentiate between possible known sources of variability and the unknown sources in scores (Haberman et al., 2008; Lee & Haberman, 2013). Their use depends on the frequency of administrations, and they are typically applied to scores by test section. Their application is irrelevant to the delivery mode of the assessment or the operational procedures for calibration, scaling, and equating. Considering the frequency of administrations, there are two types of assessments in the end: The first type of assessments involves a limited number of regular administrations per year, and the ANOVA approach used in Haberman et al. (2008) is appropriate for such assessments when there are many years for a stable operation. The second type of assessments involves continuous testing or very frequent administrations that can be examined with the family of harmonic regression approaches. Continuous testing does not have the general notation of administrations that deliver fixed forms to a group of examinees on the same date, but one may aggregate the data for a fixed testing window (e.g., weekly or biweekly) and then treat each testing window as an administration in harmonic regression. The target population of an assessment refers to all examinees who may take the test. For assessments with a moving target, empirical data from the most recent years may be used to establish their target populations (Lee et al., 2019).

Appendix A: Matrix Expressions of the Administration-Level Harmonic Regression Models for the Jackknife Results

This appendix demonstrates how the general regression model in Equation 9 connects to the administration-level models in the three approaches. Recall that the regression model in Equation 9 is given by $\mathbf{V} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$, where

- the $\mathbf{V}$ is a T-dimensional response vector $\mathbf{V}$ for the T administrations;
- the $\mathbf{Z}$ is a $T \times p$ matrix, containing a T-dimensional column of ones for the intercept and $p-1$ columns of dimension T for predictors;
- the $\boldsymbol{\gamma}$ is the p-dimensional coefficient vector; and
- the $\mathbf{e}$ is a T-dimensional error vector with elements e_t, $1 \leq t \leq T$, for administrations.

For *Average Score* (Approach 1), Equation 9 corresponds to Model 1 in Equation 2 with $p = 2K + R$, the response vector equal to $\mathbf{V} = [\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_T]'$, the $T \times p-$dimensional matrix $\mathbf{Z}$ equal to

$$\mathbf{Z} = \begin{bmatrix} 1 & \cos(2\pi d_1/L_1) & \sin(2\pi d_1/L_1) & \cdots & \cos(2\pi K d_1/L_1) & \sin(2\pi K d_1/L_1) & F_{1,1'} \cdots F_{1,R-1} \\ 1 & \cos(2\pi d_2/L_2) & \sin(2\pi d_2/L_2) & \cdots & \cos(2\pi K d_2/L_2) & \sin(2\pi K d_2/L_2) & F_{2,1'} \cdots F_{2,R-1} \\ \vdots & & & & \vdots & & \vdots \\ 1 & \cos(2\pi d_T/L_T) & \sin(2\pi d_T/L_T) & \cdots & \cos(2\pi K d_T/L_T) & \sin(2\pi K d_T/L_T) & F_{T,1'} \cdots F_{T,R-1} \end{bmatrix}, \quad \text{(A1)}$$

the coefficient vector equal to

$$\boldsymbol{\gamma} = [\mu_1, a_1, b_1, \ldots, a_K, b_K, c_1, \ldots, c_{R-1}]', \tag{A2}$$

and the error vector equal to $\mathbf{e} = [e_{1,1}, e_{1,2}, \ldots, e_{1,T}]'$. The $\mathbf{Z}$ and $\boldsymbol{\gamma}$ include terms for seasonality and regional changes of the examinees.

For *Average Residual* (Approach 2), Model 2a in Equation 5 may be expressed as Equation 9 when $p = 1$, the response vector is $\mathbf{V} = [\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_T]'$, the $\mathbf{Z} = [1, 1, \ldots, 1]'$ is a T-dimensional column of ones, the coefficient $\boldsymbol{\gamma} = \mu_3$, and the error vector $\mathbf{e} = [e_{4,1}, e_{4,2}, \ldots, e_{4,T}]'$. The matrix expression is simple for *Average Residual*, because all adjustments are accomplished in the examinee-level harmonic regression at stage 1 and no predictor is further included in the administration-level model at stage 2.

Regarding *Weighted Average* (Approach 3), Model 3 in Equation 7 is equivalent to Equation 9 given the following: $p = 2K + 1$, the response variable $\mathbf{V} = [\bar{Y}_{w,1}, \bar{Y}_{w,2}, \ldots, \bar{Y}_{w,T}]'$, the $T \times p-$dimensional matrix $\mathbf{Z}$ equal to

$$\mathbf{Z} = \begin{bmatrix} 1 & \cos(2\pi d_1/L_1) & \sin(2\pi d_1/L_1) & \cdots & \cos(2\pi K d_1/L_1) & \sin(2\pi K d_1/L_1) \\ 1 & \cos(2\pi d_2/L_2) & \sin(2\pi d_2/L_2) & \cdots & \cos(2\pi K d_2/L_2) & \sin(2\pi K d_2/L_2) \\ \vdots & & & & & \vdots \\ 1 & \cos(2\pi d_T/L_T) & \sin(2\pi d_T/L_T) & \cdots & \cos(2\pi K d_T/L_T) & \sin(2\pi K d_T/L_T) \end{bmatrix}, \tag{A3}$$

the coefficient vector

$$\boldsymbol{\gamma} = [\mu_4, a_1, b_1, \ldots, a_K, b_K], \tag{A4}$$

and the error vector $\mathbf{e} = [e_{5,1}, e_{5,2}, \ldots, e_{5,T}]'$. The $\mathbf{Z}$ and $\boldsymbol{\gamma}$ only include terms for seasonality, as the demographic changes of the examinees are accounted for through MDIA at stage 1.

Notes

[1]Lee et al. (2014) used such examinee-level predictors as gender, native country, whether the examinees took the same test before or not, and seven variables related to educational status, employment status, and experience of English learning and usage in daily life. Wei and Qu (2014) considered predictors about the examinees' education level, occupation, four variables relevant to experience of English learning and usage in daily life, how many times they have taken the same test, and reason for taking the test.

[2]For all regression models in this paper, the random errors are defined similarly with mean zero but different variances. To avoid redundancy, their definition is not repeated later.

References

Allalouf, A., Gutentag, T., & Baumer, M. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement: Issues and Practice*, *36*, 58–68.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275–285.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33–51.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika*, *75*, 438–453.

Guo, H., Robin, F., & Dorans, N. J. (2017). Detecting item drift in large-scale testing. *Journal of Educational Measurement*, *54*, 265–284.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, *12*, 971–988.

Haberman, S. J. (2010). *Limits on the accuracy of linking*. Research Report RR-10-22. Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information*. Research Memorandum RM-14-01. Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*, 254–273.

Haberman, S. J., & Dorans, N. J. (2011). *Sources of score scale inconsistency*. Research Report RR-11-10. Princeton, NJ: Educational Testing Service.

Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT® I reasoning test score conversions*. Research Report RR-08-67. Princeton, NJ: Educational Testing Service.

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy*. Research Report RR-09-39. Princeton, NJ: Educational Testing Service.

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

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

Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, *16*(3), 240–267.

Lee, Y.-H., Haberman, S. J., & Dorans, N. J. (2019). Use of adjustment by minimum discriminant information in linking constructed-response test scores in the absence of common items. *Journal of Educational Measurement*, *56*, 452–472.

Lee, Y.-H., Liu, M., & von Davier, A. A. (2014). Detection of unusual test administrations using a linear mixed effects model. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th Annual psychometric society meeting* (pp. 133–149). New York: Springer.

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.

Li, D., Li, S., & von Davier, A. A. (2011). Applying time-series analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York: Springer.

Liu, J., & Yoo, H. (2019). *Monitoring the scale stability using harmonic regression*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*(1), 46–57.

Petersen, N., Cook, L., & Stocking, M. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*(2), 137–156.

Puhan, G. (2008). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, *22*(1), 79–103.

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 RR-17-50. Princeton, NJ: Educational Testing Service.

SAS Institute Inc. (2018). *SAS/STAT® 15.1 user's guide*. Cary, NC: SAS Institute Inc.

Wei, Y., & Qu, Y. (2014). *Using multilevel analysis to monitor test performance across administrations*. Research Report RR-14-29. Princeton, NJ: Educational Testing Service.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer.

Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, *53*, 131–151.

Authors

YI-HSUAN LEE is Principal Research Scientist at Educational Testing Service, 660 Rosedale Road, MS 12T, Princeton, NJ 08541; ylee@ets.org. Her primary research interests include analysis of timing and process data, test security, quality control of assessment, item response theory, and equating and linking.

SHELBY J. HABERMAN is an independent consultant, Barak 3/1, Jerusalem 9350276, Israel; haberman.statistics@gmail.com. His primary research interests include analysis of qualitative data, sample weighting, item-response theory, equating and linking, analysis of subscores, and test security.