Research Report

Model Adequacy Checking for Applying Harmonic Regression to Assessment Quality Control

ETS RR–21-13

Jiahe Qian
Shuhong Li

December 2021

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

Model Adequacy Checking for Applying Harmonic Regression to Assessment Quality Control

Jiahe Qian & Shuhong Li

ETS, Princeton, New Jersey

In recent years, harmonic regression models have been applied to implement quality control for educational assessment data consisting of multiple administrations and displaying seasonality. As with other types of regression models, it is imperative that model adequacy checking and model fit be appropriately conducted. However, there has been no literature on how to perform a comprehensive model adequacy evaluation when applying harmonic regression models to sequential data with seasonality in the educational assessment field. This paper is intended to fill this gap with an illustration of real data from an English language assessment. Two types of cross-validation, leave-one-out and out-of-sample, were designed to measure prediction errors and check model validation. Three types of R-squared (R^2, R^2_{adj}, and R^2_{pred}) and various residual diagnostics were applied to check model adequacy and model fitting.

Keywords Sequential data with seasonality; model selection; harmonic pair index; leave-one-out cross-validation; out-of-sample cross-validation; rotated jackknife grouping

doi:10.1002/ets2.12327

Model adequacy checking is essential to assessing model fit for linear regression, logistic regression, generalized linear models, as well as harmonic regressions (HR; see Courant, 1937). For assessments with administrations (admins) across four seasons, HR models (Andrews, 2013; Lee & Haberman, 2013) are sometimes applied to conducting assessment quality control (QC) and monitoring quality of reported scores (Andrews, 2013; Lee & von Davier, 2013; Li & Qian, 2018). Although models such as ANOVA can also be used (Haberman et al., 2008), HR models typically perform better for data with many admins across seasons because the harmonic terms in HR can properly capture the seasoning variations (Lee & Haberman, 2013). Prior to the QC application in the assessment field, the HR models had found widespread applications in data with periodicity in economics, medicine, and meteorology research (Artis et al., 2007; Gaffney et al., 1993).

No existing literature has addressed how to conduct a comprehensive model adequacy evaluation when applying HRs to an educational assessment with multiple forms and periodical data. For such an assessment, if the selected HR model fits the data well, this suggests that the HR model is able to account for the expected variations in the data and the scale score trends for the assessment are stable as assessment experts expected. If there are considerable outliers (Tukey, 1977), the data can fail to confirm the HR model. The outliers are most likely due to unexpected factors that might have occurred regarding the test instrument, test administration, or the test-taker population. Therefore, for example, evaluating seasonality, which is the first step in model adequacy checking, is a prerequisite for applying HR models. Otherwise, the analysis might result in models that fail to explain the genuine relationship between the dependent and independent variables and lead to incorrect conclusions. Almost all literature in the field dwells on only this initial step (i.e., evaluating outliers and seasonality, as well as on regular model fitting examination such as model selection and checking assumptions). In terms of evaluating model adequacy for HR, some authors used different types of R-squared, root mean squared errors (RMSE) of prediction, and various residual diagnostics (Lee & Haberman, 2013; Li & Qian, 2018).

The goal of this study is to focus on exploring model adequacy checking methods for HR models with sequential data displaying seasonality. In addition to conventional techniques in checking model fitting and assumptions (Anscombe & Tukey, 1963; Cox, 2002; Ramsey, 1969) and employing graphics to analyze data periodicity and residuals (Cook, 1994; Cook & Weisberg, 1997), one primary inquiry is to investigate how to effectively conduct cross-validation, both leave-one-out and out-of-sample, for sequential data with seasonality.

Corresponding author: J. Qian, E-mail: jqian58@gmail.com

Cross-validation (Draper & Smith, 1998), a standard statistical technique for checking model adequacy, is applied for measuring prediction errors and validating model performance. Two types of cross-validation are proposed: leave-one-out cross-validation using R^2_{pred} and predicted residual error sum of squares (*PRESS*) statistics, and out-of-sample forecast using validation data to assess the model (Allen, 1974; Tarpey, 2000; Weisberg, 2014). In addition to typical indices including the three types of R-squared (R^2, R^2_{adj}, and R^2_{pred}) and RMSE of prediction (Darlington, 1968; Draper & Smith, 1998), root mean predicted residual error sum of squares (*RMPRESS*) is proposed to assess model fit and adequacy. *RMPRESS*, based on the principle of leave-one-out cross-validation, measures the average prediction errors and can be directly compared with the RMSE of prediction.

In conducting the out-of-sample cross-validation (Stone, 1974; Weisberg, 2014), data are usually partitioned into two parts, a training data set and a validation data set, and we then examine whether the models constructed from the training data can be extended to the validation data (Arlot & Celisse, 2010; Snee, 1977). For data exhibiting seasonality, we argue that random partitioning of data is not desirable for sequential data with seasonality, because it will be difficult to detect a trend among the predicted points as they blend with observed points that are subject to periodic variation. Therefore, an appropriate strategy to conduct cross-validation is to use period-based partition, that is, splitting the data into two disjoint integrated sub data sets; one set is treated as training data and the other is used for validation. Such a scheme, unlike bootstrapping design (Efron & Tibshirani, 1997), is analogous to grouped jackknifing design (Haberman et al., 2009). Further, the predicted points on the validation data can be compared with true values in assessing prediction accuracy, like the subsample-based design in survey sampling (Qian et al., 2013). Under the framework of the out-of-sample cross-validation, for evaluating prediction accuracy, a rotation design of jackknife grouping has been proposed to estimate the jackknifed variance of prediction at each point for sequential data.

The paper also addresses how to monitor model overfitting (Harrell Jr., 2001) in the event of very high R-squared values. Overfitting refers to a not-“best”-fitted HR model with a high R-squared value, making the model look unrealistically good because of limitations of sample size, the range of the predictors, and possibly the related degrees of freedom. Examples of various types of overfitting can be found in Babyak (2004). An overfitted model does not describe the genuine relationship between the dependent and independent variables. In prediction, such models would fail to conform to and extend the same shape of observed points onto future samples, thus creating significant uncertainty about the trustworthiness of the findings.

The next section covers the method used in the study, including the HR model, principle of employing harmonic terms in pairs, several proposed processes, three types of R-squared, RMSE of prediction, and RMPRESS. The Results section provides empirical graphs, including the plots of mean scores across four seasons for all four skills in the assessment and the plots of residuals versus the predicted values of the HR models. The model adequacy checking focuses on applying cross-validation to model selection and model specification. The final section offers a summary with major conclusions.

Method

Data

The real data analyzed in the study contain the admin mean scale scores of four different skills (Skill A, Skill B, Skill C, and Skill D) across four seasons in six consecutive years of a large-scale English language assessment. Table 1 presents the summary statistics for the data. For reliable results, the analysis includes only admins with sample sizes greater than 3,000; the total number of included admins is 498. To thoroughly assess the regional effects on the test-taker performance of the English language assessment, the data sets include 19 dummy region variables, which can be found in the first column of Table 2.

Harmonic Regression Model

A general HR model (Courant, 1937; Lee & Haberman, 2013; Li & Qian, 2018) is constructed for the data with expected factors that cause measurement variability for the assessment. These factors include three main types of predictors: dummy variables for year effect, dummy variables for regional effects, and harmonic terms for the sum of sine and cosine functions that reflect seasonal trends. The dummy variable for year effect has two levels: 1 if the observation is of the year and 0 for otherwise. A significant year effect indicates the existence of trend changes. In addition, because we have preknowledge about the assessment that test takers’ performance tends to vary across different regions, incorporating regional effects

Table 1 The Means and Standard Deviations of the Admin Mean Scale Scores for the Four Skills in the Six Years' Data

Section statistic	Year	Statistic mean	Statistic *SD*
Skill A	Year 1	14.745	0.895
	Year 2	15.099	0.736
	Year 3	15.123	0.794
	Year 4	15.302	0.864
	Year 5	15.483	0.908
	Year 6	15.854	0.756
Skill B	Year 1	14.560	0.963
	Year 2	14.898	0.889
	Year 3	14.823	1.143
	Year 4	15.005	1.256
	Year 5	15.157	1.395
	Year 6	15.432	1.073
Skill C	Year 1	15.284	0.932
	Year 2	15.360	0.936
	Year 3	15.317	1.003
	Year 4	15.364	1.172
	Year 5	15.358	1.277
	Year 6	15.462	0.880
Skill D	Year 1	15.616	0.537
	Year 2	15.786	0.459
	Year 3	15.379	0.564
	Year 4	15.604	0.629
	Year 5	15.799	0.774
	Year 6	15.873	0.603

into the model is expected to improve model fit and differentiating power (Lee & Haberman, 2013; Wang et al., 2018). Note that in fitting HR, depending on the characteristics of data and specific research topic(s) of interest, other predictors can be included as necessary.

In this study, a full HR model is

$$y_t = \mu + \sum_{i=1}^{H} \left[\alpha_i \cos\left(\frac{2i\pi d_t}{T_t}\right) + \beta_i \sin\left(\frac{2i\pi d_t}{T_t}\right) \right] + \sum_{j=2}^{6} \gamma_j x_{jt} + \sum_{k=1}^{18} \delta_k f_{kt} + e_t, \tag{1}$$

where y_t is the admin mean score at time point[1] t and $e_t \underset{iid}{\sim} N\left(0, \sigma^2\right)$ is the residual term;

$$\sum_{i=1}^{H} \left[\alpha_i \cos\left(\frac{2i\pi d_t}{T_t}\right) + \beta_i \sin\left(\frac{2i\pi d_t}{T_t}\right) \right]$$

is the sum of the pairs of *harmonic terms*, $\cos(\cdot)$ and $\sin(\cdot)$, with coefficients α_i and β_i (harmonic pair index $H = 1, 2, 3$, etc., and a *full H* refers to the highest H of an HR) in Equation 1, d_t is the number of days elapsed for the admin at point t since the beginning of the year, and T_t is the total number of days in the year;

$$\sum_{j=2}^{6} \gamma_j x_{jt}$$

is the sum of the *year-effect* terms, with coefficients γ_j and dummy year variables x_{jt} (the year index $j = 2, \ldots, 6$) for admin at point t;

$$\sum_{k=1}^{18} \delta_k f_{kt}$$

is the sum of the *region-effect* terms, with coefficients δ_k and dummy region variables f_{kt} (the region index $k = 1, 2, \ldots, 18$) are expressed in proportion of the test takers in region k for admin at point t. The names of the 19 region variables can be found in the first column of Table 2 except for f_{19t}, Mid-East abroad. The dummy variables f_{19t}, as well as x_{1t}, are dropped from the equation to avoid perfect multicollinearity in the model (Draper & Smith, 1998).

Table 2 The Parameter Estimates of the Harmonic Regression Models of Admin Means With Terms Retained by the Backward Selection Procedure on Year 1 – Year 6 Full Data and Year 1 – Year 5 Training Data

	Parameter estimates on Year 1 – Year 6				Parameter estimates on Year 1 – Year 5			
Variables	Skill A	Skill B	Skill C	Skill D	Skill A	Skill B	Skill C	Skill D
Intercept	8.705[a]	10.894[a]	18.694[a]	18.975[a]	8.764[a]	10.632[a]	18.682[a]	19.474[a]
Year 2	0.287[a]	0.415[a]	0.137[a]	0.061	0.279[a]	0.413[a]	0.121[a]	0.043
Year 3	0.389[a]	0.445[a]	0.174[a]	-0.294[a]	0.378[a]	0.443[a]	0.152[a]	−0.307[a]
Year 4	0.534[a]	0.641[a]	0.192[a]	−0.159	0.545[a]	0.652[a]	0.179[a]	−0.176[a]
Year 5	0.685[a]	0.873[a]	0.286[a]	0.068	0.702[a]	0.884[a]	0.274[a]	0.057
Year 6	1.043[a]	1.220[a]	0.508[a]	0.204				
β_1 sin	−0.717[a]	−0.689[a]	−0.275[a]	−0.337[a]	−0.696[a]	−0.663[a]	−0.245[a]	−0.315[a]
α_1 cos	0.063	0.004	0.023	0.118[a]	0.083	0.004	0.045	0.129[a]
β_2 sin	−0.002	0.046	0.082[a]		−0.011	0.046	0.075[a]	
α_2 cos	−0.221[a]	−0.142[a]	−0.055[a]		−0.238[a]	−0.154[a]	−0.060[a]	
Africa			5.892				5.888	
Africa abrd.			−24.442				−24.771	
Ameri.	10.152[a]	8.490[a]			10.146[a]	8.804[a]	0.604	
Ameri. abrd.	15.158	10.109			17.142	11.885		
Asia 1	9.440[a]	5.411			8.475[a]	5.173		
Asia 1 abrd.	31.399	27.295		−17.646	24.618	24.595		−17.084
Asia 2	11.200[a]	6.872[a]		1.309[a]	11.171[a]	7.102[a]		0.796
Asia 2 abrd.	10.154[a]	9.734[a]	3.685[a]	4.479[a]	9.642[a]	9.863[a]	3.572[a]	3.906[a]
Eng. Spk Ctry[#]								
Europe	14.645[a]	13.363[a]	4.212[a]	4.319[a]	14.567[a]	13.639[a]	4.179[a]	3.880[a]
Europe abrd.			9.330				7.912	
Asia 3	12.499[a]	11.051[a]	3.915[a]	4.827[a]	12.528[a]	11.353[a]	3.928[a]	4.213[a]
Asia 3 abrd.								−10.618[a]
Asia 4	8.078[a]	6.301[a]	−1.852[a]		8.238[a]	7.238[a]	−1.467	
Asia 4 abrd.								
Asia 5	13.860[a]	10.350[a]	2.122[a]	3.238[a]	13.814[a]	10.501[a]	2.116[a]	2.217[a]
Asia 5 abrd.	18.477[a]	15.693	8.398[a]	11.082	22.089[a]	18.051	8.548[a]	10.893
Mid-E!	11.876[a]	10.487[a]	2.754[a]		12.324[a]	10.832[a]	2.502[a]	

Note. The blank cells are the terms not retained by the backward model selection. Ameri. = Americas; abrd. = abroad (i.e., taking the test outside of their home country); Eng. Spk Ctry = English-speaking country, including Austria, Canada, New Zealand, and Singapore; Mid-E! indicates that the variable Mid-E abrd (the counterpart of Mid-E) was not used in modeling. [a]Significance in *t*-tests with the Bonferroni correction based on selected models.

Principle: Employing Harmonic Terms in Pairs

As a periodic function, the signal component in sequential data with seasonality can be analyzed by a Fourier series composed of terms of harmonically related sinusoids with a phase angle

$$A_i \cos\left(\frac{2i\pi d_t}{T_t} - \varphi_i\right), \tag{2}$$

where A_i is amplitude and φ_i is the phase angle for $i = 1, 2, \ldots H$ (Tolstov, 1976). The phase angle φ_i is a variate with nonzero expectation for educational assessments administered across seasons.

The harmonic terms in Equation 1 are always used in pairs because term *i* in a Fourier series within amplitude-phase form can be expressed as

$$A_i \cos\left(\frac{2i\pi d_t}{T_t} - \varphi_i\right) = A_i \cos\left(\varphi_i\right) \cos\left(\frac{2i\pi d_t}{T_t}\right) + A_i \sin\left(\varphi_i\right) \sin\left(\frac{2i\pi d_t}{T_t}\right), \tag{3}$$

where $A_i = \sqrt{\alpha_i^2 + \beta_i^2}$ and the phase angle $\varphi_i = \arctan\left(\frac{\beta_i}{\alpha_i}\right)$ (Tolstov, 1976). Equation 3 can be derived from a trigonometric identity. Because

$$\cos\left[\arctan\left(\frac{\beta_i}{\alpha_i}\right)\right] = \frac{1}{\sqrt{\left(\frac{\beta_i}{\alpha_i}\right)^2 + 1}} = \frac{\alpha_i}{\sqrt{\beta_i^2 + \alpha_i^2}},$$

and, similarly,

$$\sin\left[\arctan\left(\frac{\beta_i}{\alpha_i}\right)\right] = \frac{\beta_i}{\sqrt{\beta_i^2 + \alpha_i^2}},$$

the coefficients $A_i \cos\left(\varphi_i\right) = \alpha_i$ and $A_i \sin\left(\varphi_i\right) = \beta_i$. Thus the right side of Equation 3 can be expressed as

$$\alpha_i \cos\left(\frac{2i\pi d_t}{T_t}\right) + \beta_i \sin\left(\frac{2i\pi d_t}{T_t}\right). \tag{4}$$

Consequently, a periodic function in amplitude-phase form equals the harmonic terms with sine and cosine pair in Equation 4, that is,

$$A_i \cos\left(\frac{2i\pi d_t}{T_t} - \varphi_i\right) = \alpha_i \cos\left(\frac{2i\pi d_t}{T_t}\right) + \beta_i \sin\left(\frac{2i\pi d_t}{T_t}\right).$$

Because the cosine term is nonlinear in Equation 2, the significance test of the sinusoid estimates in Equation 2 is complex. Therefore, instead of using sinusoids, the pairs of harmonic terms as in Equation 4, serving as a Fourier basis, are always used in HR (Courant, 1937). The principle explains why harmonic terms should be used in pairs when we fit HR models.

The Process to Determine the Harmonic Term Pairs

To determine the harmonic term pairs in Equation 1 means to identify the *full H* (i.e., the highest H of an HR). For example, in Table 2, we have *full* $H = 1$ for Skill D in Equation 1 on the Year 1 – Year 6 full data and *full* $H = 2$ for Skill A, Skill B, and Skill C. When $H = 2$, the two included pairs of the harmonic terms for each of the skills are $\alpha_1 \cos\left(\frac{2\pi d_t}{T_t}\right)$ and $\beta_1 \sin\left(\frac{2\pi d_t}{T_t}\right)$ for $i = 1$, and $\alpha_2 \cos\left(\frac{4\pi d_t}{T_t}\right)$ and $\beta_2 \sin\left(\frac{4\pi d_t}{T_t}\right)$ for $i = 2$.

The inclusion of harmonic pairs is based on the significance test of the parameter estimates of the harmonic term pairs in Equation 1. A stepwise test process can be used to determine the *full H*. First, for $H = 1$, construct the full HR with one pair of harmonic terms and check the significance of the parameter estimates $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ of the first harmonic pair. If neither is significant, the model is a regular regression and the process ends; otherwise, if either of the parameter estimates is significant, it suggests $H = 1$ holds, and the first pair of harmonic terms is included in Equation 1. Then, for $H = 2$, construct the full HR with two pairs of harmonic terms and examine the significance of the parameter estimates $\widehat{\alpha}_2$ and $\widehat{\beta}_2$. If none of the parameter estimates $\widehat{\alpha}_2$ and $\widehat{\beta}_2$ are significant, it suggests *full* $H = 1$ and the process ends; otherwise, if either is significant, $H = 2$, and the second pair of harmonic terms is also included in Equation 1. Continue the process for each of the next H until it ends. Based on empirical outcomes in literature (Andrews, 2013; Lee & Haberman, 2013; Lee & von Davier, 2013; Li & Qian, 2018), the *full H* is usually no larger than 3 in applying HR to assessment QC. The detailed application process is illustrated with real data under the Model Selection subhead in the Results section.

The Process to Conduct Out-of-Sample Cross-Validation

As mentioned in the introduction, out-of-sample cross-validation is conducted to validate whether the models constructed from the training data can be extended to the validation data. For sequential data with seasonality, the conventional scheme of dropping a random sample, as in a bootstrapping design, is not desirable because a sampling-based partition does not allow us to detect trends in the predicted admin points, as they blend with observed points that are subject to periodic variations. Instead, we propose dropping an integrated chunk of sample from the data as a basis upon which to assess the prediction error of the model. This is similar to a grouped jackknifing design. In this analysis, the training data consist of Year 1 to Year 5 admins, and the validation data consist of Year 6 admins. There are alternative ways to conduct the data partition. For example, the training data contain Year 1 to Year 4 admins and the validation data comprise Year 5 and Year 6 admins (see more discussion in the Model Adequacy Checking: Out-of-Sample Cross-Validation section of the Results).

Because the year effects are significantly nonzero for all four skills, to improve prediction accuracy, the predicted year effects, the term $\widehat{\gamma}_6 x_{6t}$ for admin at t in Year 6 in Equation 1, are included in the prediction. In this study, the technique of hot deck imputation (Little & Rubin, 2002) was used to estimate the coefficient $\widehat{\gamma}_6$; specifically, it took the same value as

$\hat{\gamma}_5$ in predicting the means because the predicted means estimated by the validation data of Year 6 had the same tendency patterns as those for the training data of Year 5, with small prediction errors. As pointed out by a reviewer, the hot deck prediction could be confounded by the year drift effect and other predictors. To further improve prediction accuracy, we might need to use enhanced statistical techniques such as autoregressive integrated moving average (ARIMA) for time series (Guo et al., 2017; Hillmer & Tiao, 1982).

Three Types of R-Squared, Root Mean Squared Error of Prediction, and Root Mean Predicted Residual Error Sum of Squares

Five statistical indices, including three types of R-squared, RMSE of prediction, and RMPRESS, are used to examine model fit as well as in model selection. Two statistical indices, R^2_{pred} and *PRESS*, are used for the leave-one-out cross-validation analysis.

Let $\bar{y}_.$ be the sample average. Let $\hat{y}_t$ be the predicted value for admin at point t and $\hat{y}_{(t)}$ be the predicted value yielded by the data with admin at point t dropped. For a sample of size n, define $SST = \sum_{t=1}^{n} \left(y_t - \bar{y}_.\right)^2$ and $SSE = \sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2$. The RMSE of prediction for a model is defined as the square root of the mean of the squared differences between the true criterion value for each record and the predicted value yielded by the regression:

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{5}$$

(Darlington, 1968).

The R-squared is as

$$R^2 = 1 - \frac{SSE}{SST}. \tag{6}$$

The adjusted R^2 is defined by making adjustments for the corresponding degrees of freedom to the terms of R^2:

$$R^2_{\text{adj}} = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-g}\right), \tag{7}$$

where g is the number of independent variables (Draper & Smith, 1998). R^2_{adj} measures the percentage of variation explained only by the independent variables.

The *PRESS* residual is defined as

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}\right)^2 \tag{8}$$

(Allen, 1974; Draper & Smith, 1998; Pierce & Schafer, 1986; Tarpey, 2000). The *PRESS*, a form of leave-one-out cross-validation, is a summary measure of model fit based on $\hat{y}_{(i)}$ $(= 1, \ldots\ n)$, in which the observation i is not included to estimate the model.

The predicted R^2 is defined as

$$R^2_{\text{pred}} = 1 - \frac{PRESS}{SST}. \tag{9}$$

The R^2_{pred} measures the prediction power based on leave-one-out cross-validation. Compared with R^2 and R^2_{adj}, a low value of R^2_{pred} indicates poor prediction of the model and a possibility of model overfitting, in particular if the difference between R^2_{pred} and R^2_{adj} is greater than 0.20 (McClellan & Staiger, 2000). Therefore, R^2_{pred} is an effective tool for model adequacy checking.

Lastly, the *RMPRESS* is defined as

$$RMPRESS = \sqrt{\frac{PRESS}{n}} \tag{10}$$

RMPRESS quantifies the average prediction error and indicates the prediction validity for a regression analysis. Although RMSE also assesses the average prediction error of a model, *RMPRESS* is estimated by the leave-one-out cross-validation approach and provides more adequate prediction error. See the Model Adequacy Checking: Model Overfitting section in Results for the specific application process. Table B1 in Appendix B presents the summary and comparison of all foregoing statistics used to examine model fit.

Results

Model Selection

As noted in the introduction, the preliminary task for HR model selection was to verify whether data displayed periodicity across seasons and whether HR was the appropriate model. Figures 1 and 2 present the plots of mean scale scores of the four skills of all admins across 6 years. The dark points in the plots are the moving averages of the mean scale scores of every six adjacent administrations. The periodic trends suggested a clear pattern of seasonality. In Figures 1 and 2, a 3*SD band, indicated by two dotted lines, was formed above and below the overall mean $\overline{y}_{.}$ of each skill. There were no outlier admins in the trend because all the admin points fell within the band (Tukey, 1977). The empirical evidence described above supported the usage of the data in HR modeling.

Model selection consisted of two stages. At the first stage, we used significance test to determine the harmonic pair index H in Equation 1. For the model of Skill A on the Year 1 – Year 6 data, when $H = 1$, the coefficients $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ for the trigonometric terms of cosine and sine are -0.741^* and 0.122, respectively. Here, symbol * in -0.741^* indicates significance in t-tests with the Bonferroni correction. Thus, the first pair of harmonic terms $i = 1$ was included in Equation 1. Next, when $H = 2$ for the model of Skill A, the coefficients $\widehat{\alpha}_2$ and $\widehat{\beta}_2$, in Table 2, are -0.002 and -0.221^*, respectively. Because $\widehat{\beta}_2$ was significant, the pair of $i = 2$ was also included in the equation. However, when $H = 3$, results (not reported in Table 2) indicated that the coefficients $\widehat{\alpha}_3$ and $\widehat{\beta}_3$, -0.002 and -0.02, were both not significant, so the process ended. $H = 2$ is called *full H* of the model for Skill A, that is, the model included harmonic term pair $\alpha_2 \cos\left(\frac{4\pi d_t}{T_t}\right)$ and $\beta_2 \sin\left(\frac{4\pi d_t}{T_t}\right)$ and the pair $\alpha_1 \cos\left(\frac{2\pi d_t}{T_t}\right)$ and $\beta_1 \sin\left(\frac{2\pi d_t}{T_t}\right)$. The models for Skill B and Skill C had a similar *full H* $= 2$. The Skill D model had the *full H* $= 1$, because $\widehat{\beta}_1$ (for sine) in the pair of cosine and sine terms, 0.002 and -0.221^*, was significant when $H = 1$, but $\widehat{\alpha}_2$ and $\widehat{\beta}_2$, -0.002 and -0.01, were both not significant when $H = 2$. By the principle of employing harmonic terms in pairs, the Skill D model had the *full H* $= 1$.

At the second stage, after the harmonic term pairs were determined, we conducted the process to select parsimonious models from models with no autocorrelation measured by the correlations between adjacent cases in the residuals (Lee & Haberman, 2013). In this process, the Durbin – Watson d (Durbin & Watson, 1971) was used to detect autocorrelation in the analysis. In Table 3, all the Durbin – Watson d statistics for the HR models were close to 2 (well above 1 and below 4), suggesting that these models showed no apparent autocorrelation (Draper & Smith, 1998).

Subsequently, given the included harmonic term pairs, backward selection procedure (Draper & Smith, 1998; Hocking, 1976; Judge et al., 1980) was used to select the region variables in an F test with a significance alpha entry level of .25 and a staying alpha level of .15. Two alpha levels were chosen by the rule of thumb in empirical regression analyses (SAS Institute, 2008). Note that those dropped variables accounted for some of the blank cells in Table 2. We then conducted the significance tests to the list of variables retained from the first stage. The significance of the variables was checked using the t – test with the Bonferroni correction (Lee & Haberman, 2013), by setting the significance cutoff at α/n^*, where $\alpha = .05$ and n^* referred to the number of retained variables in the selected models. The significance test of a shorter list of retained variables with the Bonferroni correction was less conservative in comparison with the previous analysis (Li & Qian, 2018). Thus, we were able to select an adequate model and detect the significant predictors.

Table 2 presents the parameter estimates of the parsimonious HR models of the admin means for each skill based on the full data of Year 1 to Year 6 and the training data of Year 1 to Year 5. Note that the Year 1 to Year 5 data were called training data because the data were used to build prediction models in the out-of-sample cross-validation; moreover, the comparison between the results from the model derived from the full data and those from the training data showed that the HR models were stable and effective for QC purposes.

Table 3 provides the summary statistics for each skill, including the estimates of Akaike's information criterion (AIC; see Akaike, 1973), RMSE of prediction, RMPRESS, R^2, R^2_{adj}, and R^2_{pred} of the models for the full data and the training data, respectively. The RMSE values suggested that the models demonstrated reasonable model fit because, except for Skill C, RMSE estimates were between .43 and .50. The R^2 estimates ranged between .58 and .87, except for Skill C, which also suggested reasonable model fit. The implications of a large R^2 estimate for Skill C will be discussed in detail in the Model Adequacy Checking: Model Overfitting section.

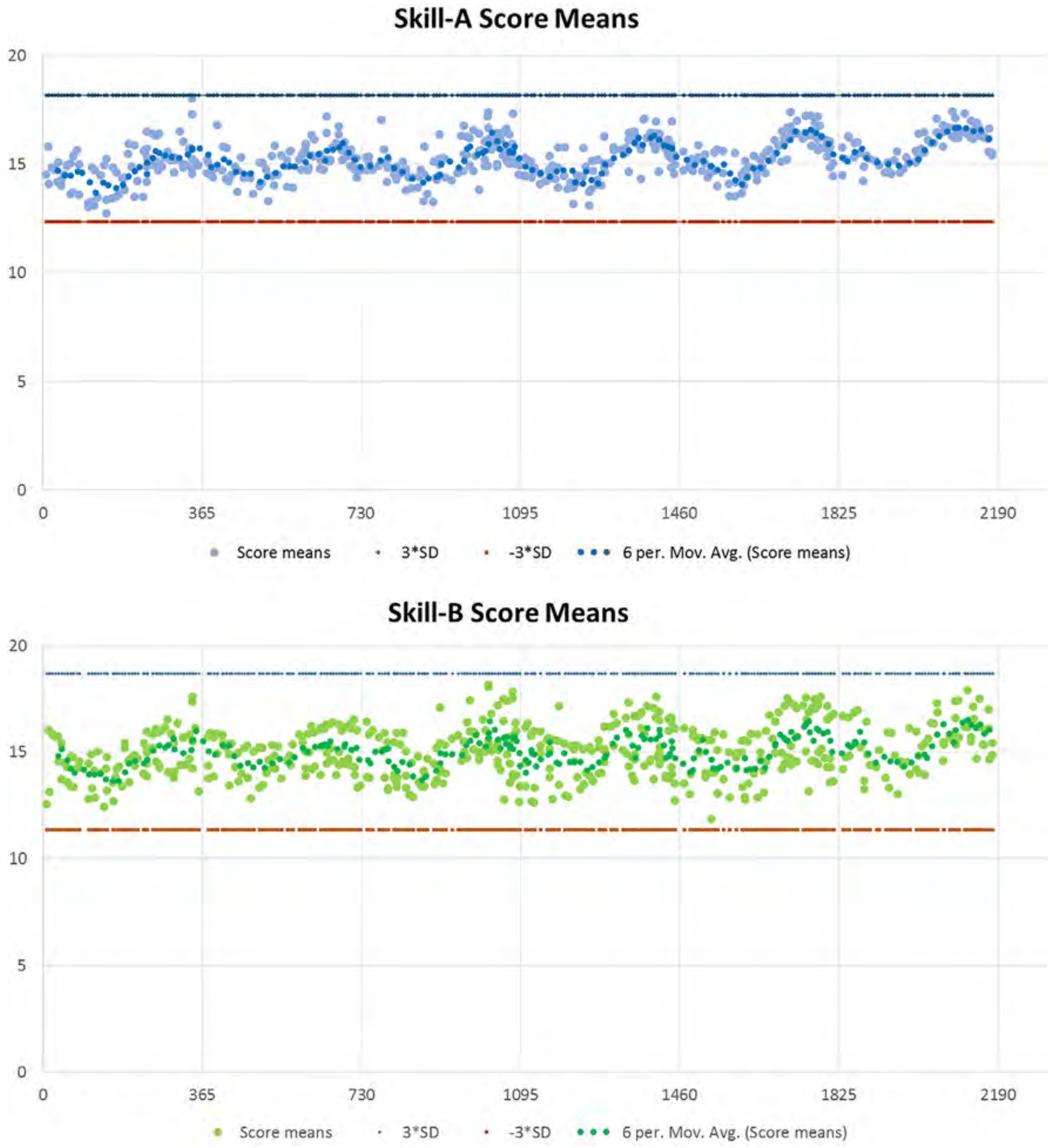

Figure 1 Mean scale scores of each test administration for all 6 years' data with moving averages of the mean scores of every six administrations (in dark points) and 3*SD band for the overall mean. Note that the Excel software does not create a moving average for every admin; it only prints disconnected dots. The test date *t* starts from the beginning time point of Year 1; there are in total 2,191 days (2,191 = 365 × 6 + 1) across Year 1 – Year 6 including 1 day from a leap year (2016).

Residual Plots

The residual plots were used to check the HR model assumptions. The regular diagnostic methods of linear regression can be found in Draper and Smith (1998). To facilitate readability, main processes and methods used in checking major regression assumptions are listed in Table B3 in Appendix B.

Figure 3 contains two types of residual plots. The first row presents the plots for the residuals versus the predicted values. In these plots, the residuals had zero means across the predicted values, and the random dispersion around the horizontal axis had no linear or seasonal trends. The second row provides the normal QQ plots for the residuals. The centered fit in the QQ plots indicated normality of the residuals. The residuals with random dispersion were uncorrelated, which was

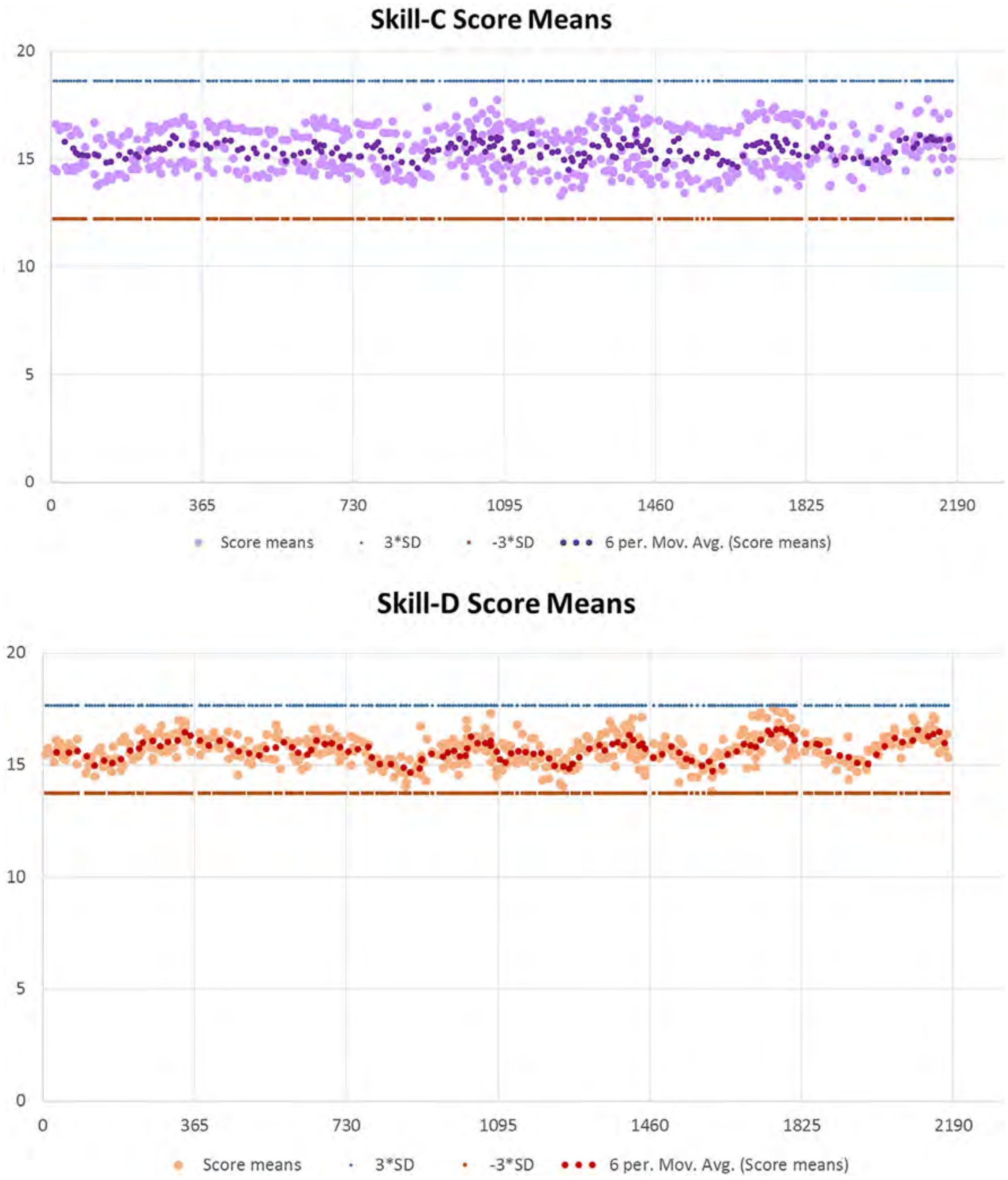

Figure 2 Mean scale scores of each test administration for all 6 years' data with moving averages of the mean scores of every six administrations (in dark points) and 3*SD band for the overall mean. Note that the Excel software does not create a moving average for every admin; it only prints disconnected dots. The test date t starts from the beginning time point of Year 1; there are in total 2,191 days ($2{,}191 = 365 \times 6 + 1$) across Year 1 – Year 6 including 1 day from a leap year (2016).

consistent with the residual assumption that $e_t \underset{iid}{\sim} N\left(0, \sigma^2\right)$. These plots confirmed that, after the harmonic term pairs had been selected, no seasonality or periodicity was observed in the remaining residuals. The Skill C plot appeared to have two scattered clusters, which might be explained by the demographic feature region of the test takers. This is consistent with our preknowledge about the assessment as well as the patterns shown in Figures 1 and 2.

Figure 4 contains the plots of residuals against region variables with the horizontal axis representing the region proportions. For example, the plots for Europe are in the second row. In alignment with the findings above, the plots, excluding

Table 3 Comparison of Akaike's Information Criterion,[2] Root Mean Squared Error of Prediction, R^2, Adjusted R^2, and Predicted R^2 for the Full Data of Year 1 – Year 6 and the Training Data of Year 1 – Year 5

Section	AIC	RMSE	RMPRESS	R^2	Adj R^2	Pred R^2	Durbin – Watson d[a]
Selected model on Year 1 – Year 6 data							
Skill A	−707.01	0.478	0.497	0.724	0.709	0.685	1.890
Skill B	−792.74	0.439	0.453	0.866	0.858	0.849	1.950
Skill C	−1, 374.84	0.245	0.254	0.949	0.946	0.941	1.742
Skill D	−844.28	0.418	0.429	0.578	0.556	0.531	1.740
Selected model on Year 1 – Year 5 data							
Skill A	−601.15	0.486	0.504	0.707	0.703	0.688	1.857
Skill B	−666.53	0.452	0.463	0.859	0.858	0.851	1.966
Skill C	−1, 191.48	0.246	0.252	0.951	0.950	0.948	1.790
Skill D	−738.99	0.417	0.424	0.583	0.576	0.560	1.749

Note. The selected models can be found in Table 2. [a] All the Durbin – Watson d are close to 2 (well above 1 and below 4); these models show no apparent autocorrelation.

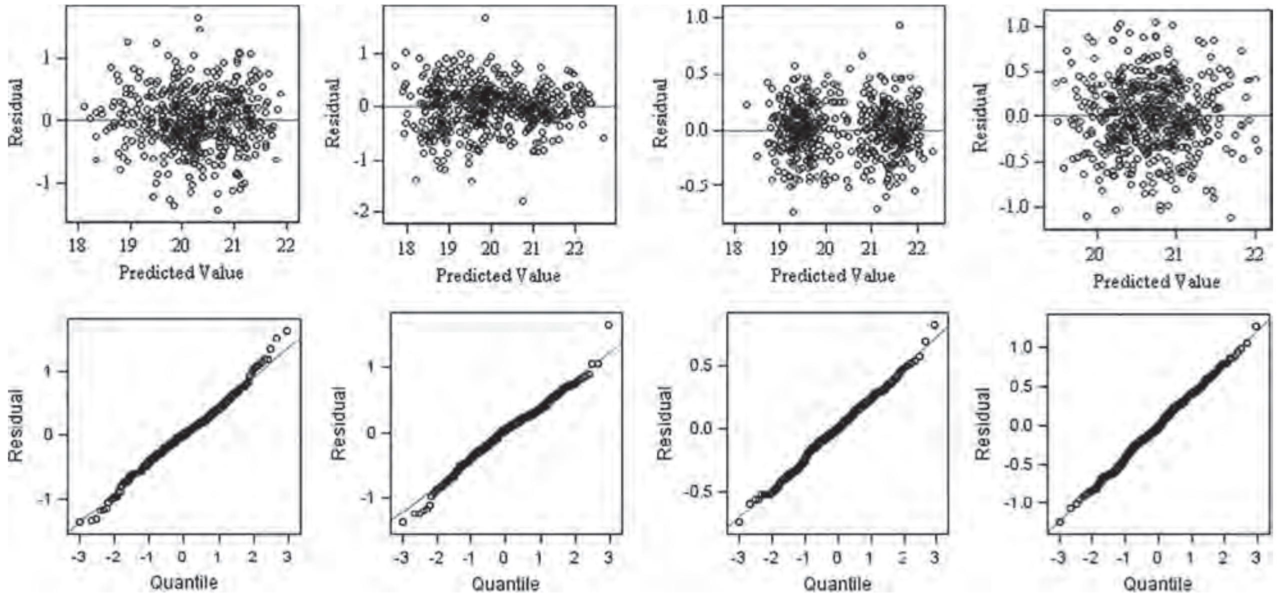

Figure 3 Residuals versus the predicted values and normal QQ plots for the residuals of the harmonic regression models (left to right: Skill A, Skill B, Skill C, and Skill D; first row: residuals vs. the predicted values; second row: normal QQ plots for the residuals).

zero points, showed no sign of obvious explanatory patterns. Note that the points clustered at zero were the admins without European test takers. The residual plots for other region variables showed similar patterns.

Model Adequacy Checking: Out-of-Sample Cross-Validation

In addition to the mean scale scores of each admin across 6 years, Figures 5 and 6 present the plots of the predicted means yielded by the training data (Year 1 to Year 5) and the validation data (Year 6), both in dark points but in different colors. Those dark points in the left segment are yielded by the model from the training data across Year 1 to Year 5; the dark points in the right segment, represented in a different color scheme, are the predicted means yielded from the validation data of Year 6. The validation data were used to assess the model yielded by the training data. Because test takers are known to exhibit different performance patterns across regional groups for this test, mainly West versus East, the predicted admin means of Skill B and Skill C, in Figures 5 and 6, respectively, demonstrated such a difference in having two distinct dark point curves across 6 years; this is also exhibited by their residual plots in Figure 3.

Because the year effects were nonzero for the training data and significant for Skill A through Skill C and part of the Year 1 – Year 5 for Skill D (see Table 2), the predicted year effects $\widehat{\gamma}_6 x_{6t}$ were added to the means of Year 6 in prediction.

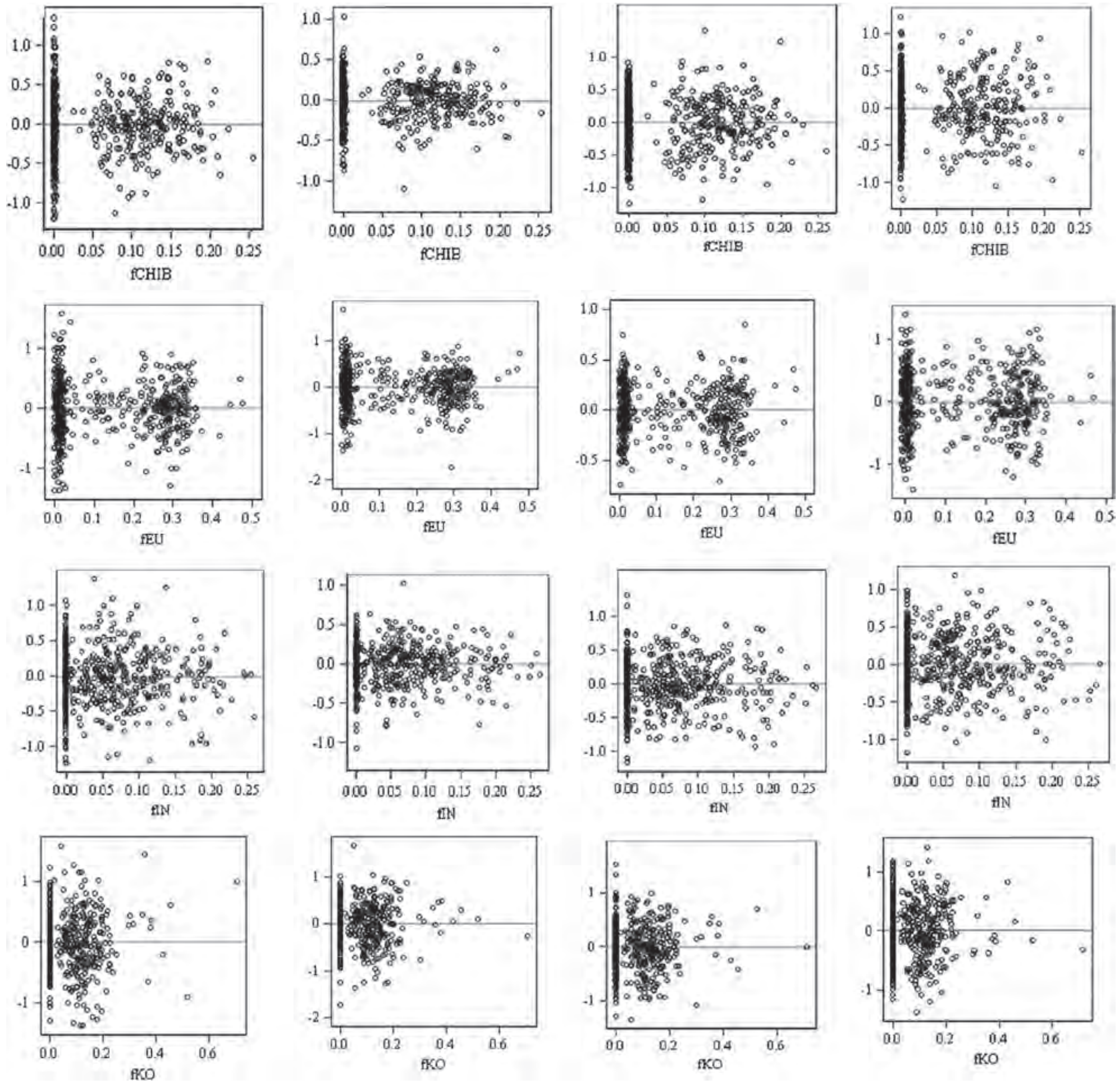

Figure 4 Residuals versus proportions of Asia2 abroad (with label fCHIB in plot), Europe (with label fEU in plot), Asia3 (with label fIN in plot), and Asia5 (with label fKO in plot) test takers for the selected models of admin means (left to right: Skill A, Skill B, Skill C and Skill D; first row: Asia2 abroad test takers; second row: Europe test takers; third row: Asia3 test takers; fourth row: Asia5 test takers).

Following the hot deck imputation (Little & Rubin, 2002), the coefficient $\hat{\gamma}_6$ took the same value as $\hat{\gamma}_5$; for example, for Skill A, the imputed value of $\hat{\gamma}_6$ was .702 for the year effects $\hat{\gamma}_6 x_{6t}$ in prediction. According to Table 1, the average means for Skill A and Skill B of Year 6 increased by .37 and .27 from those of Year 5, respectively. If the year effect term $\hat{\gamma}_6 x_{6t}$ had not been included, as shown in Figures 7 and 8, the predicted means for Year 6 validation data would have been slightly lower than their empirical counterparts, in particular for Skill A and Skill B, as in Figures 5 and 6. Obviously, the predicted means for Year 6 validation data in Figures 7 and 8 appear inferior and deviate more than their empirical counterparts in Figures 5 and 6, in particular for Skill A and Skill B. The imputation technique, as mentioned before, is based on hot deck imputation for surveys with missing responses (Little & Rubin, 2002); the predicted means with included year effects had prediction errors less than 1 *SD*. Because the effect of $\hat{\gamma}_6 x_{6t}$ on prediction could be confounded by the year drift effect and other predictors, ARIMA can be used to improve prediction accuracy. This is a topic for future research.

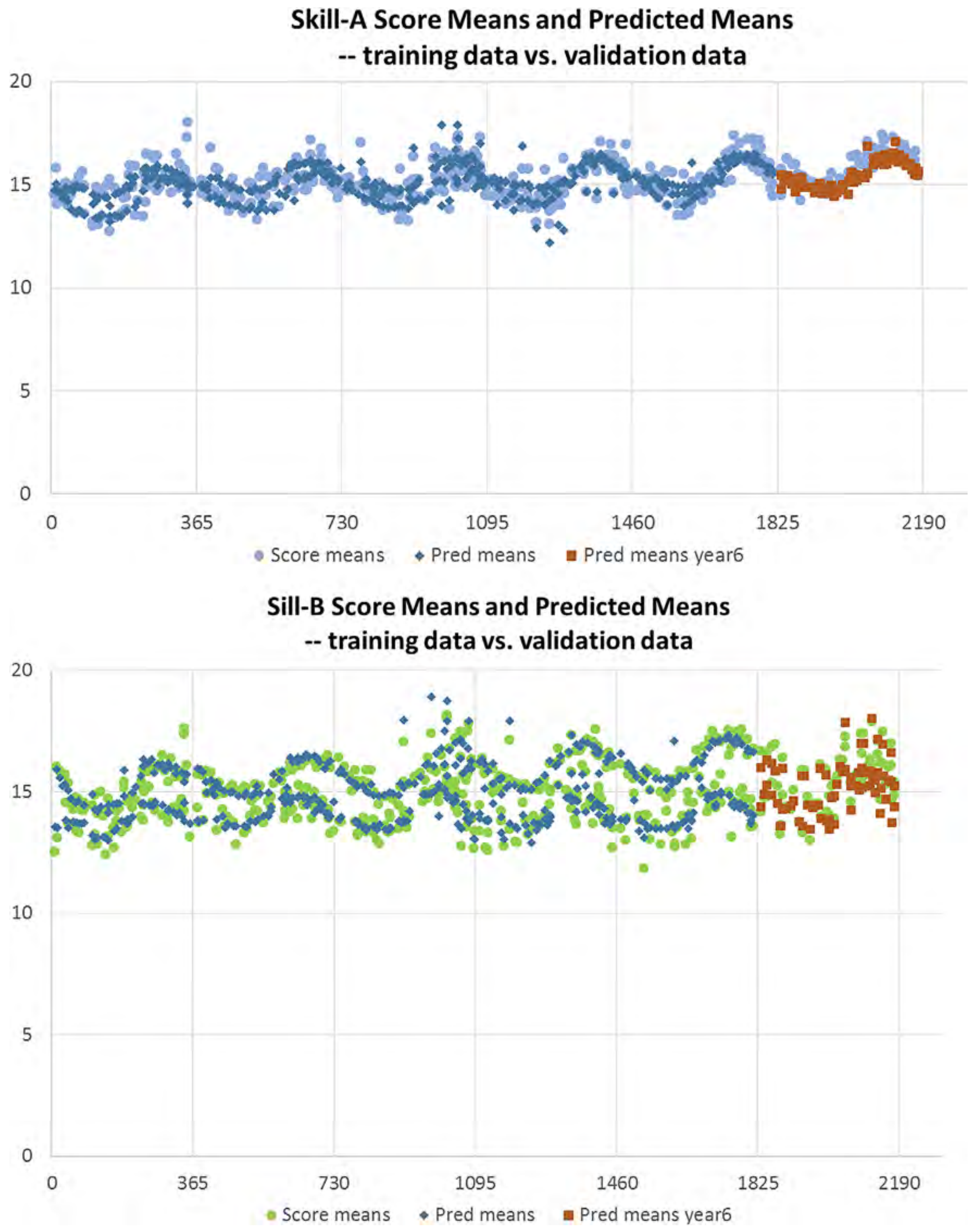

Figure 5 Empirical mean scale scores of each test administration for all 6 years' data for Skills A and B, with dark points representing the predicted means yielded by the training data (Year 1 to Year 5) and the validation data (Year 6); the prediction for Year 6 included the predicted year effects in the out-of-sample cross-validation. The test date t starts from the beginning time point of Year 1; there are in total 2,191 days ($2191 = 365 \times 6+1$) across Year 1 – Year 6 including 1 day from a leap year (2016).

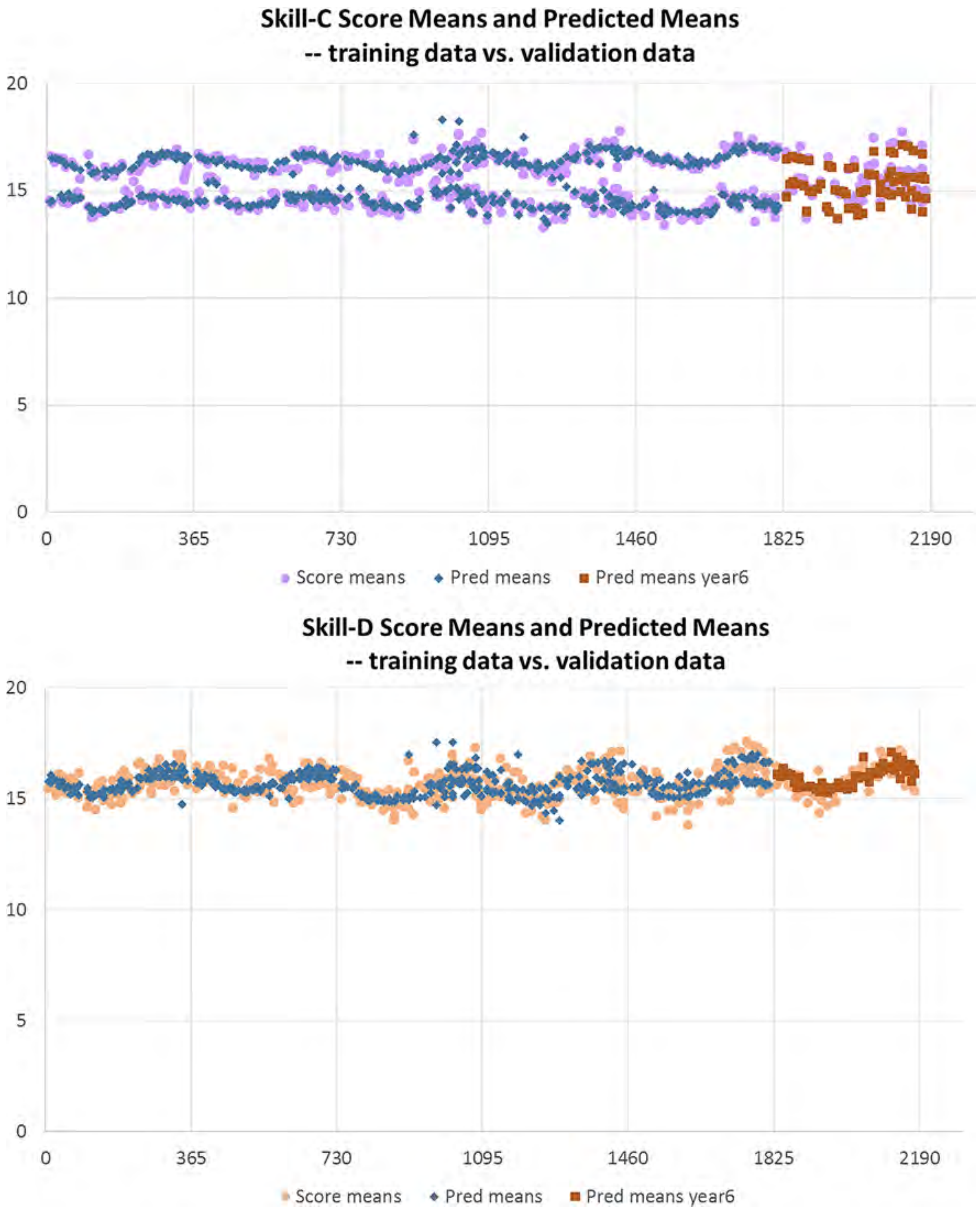

Figure 6 Empirical mean scale scores of each test administration for all 6 years' data for Skills C and D, with dark points representing the predicted means yielded by the training data (Year 1 to Year 5) and the validation data (Year 6); the prediction for Year 6 included the predicted year effects in the out-of-sample cross-validation. The test date *t* starts from the beginning time point of Year 1; there are in total 2,191 days (2,191 = 365 × 6 + 1) across Year 1 – Year 6 including 1 day from a leap year (2016).

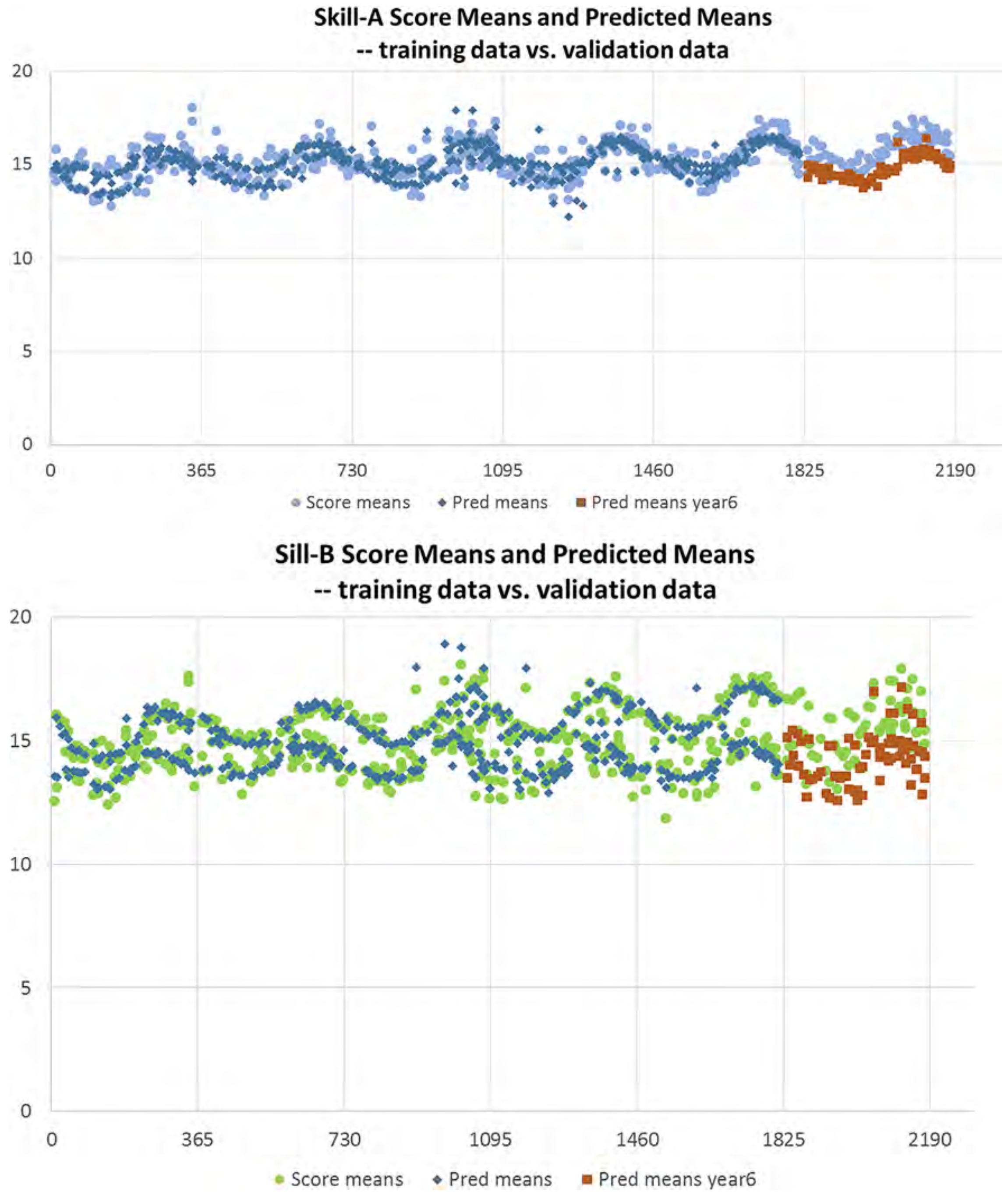

Figure 7 Empirical mean scale scores of each test administration for all 6 years' data for Skills A and B, with dark points representing the predicted means yielded by the training data (Year 1 to Year 5) and the validation data (Year 6); the prediction for Year 6 did not include the predicted year effects in the out-of-sample cross-validation. The test date t starts from the beginning time point of Year 1; there are in total 2,191 days ($2{,}191 = 365 \times 6 + 1$) across Year 1 – Year 6 including 1 day from a leap year (2016).

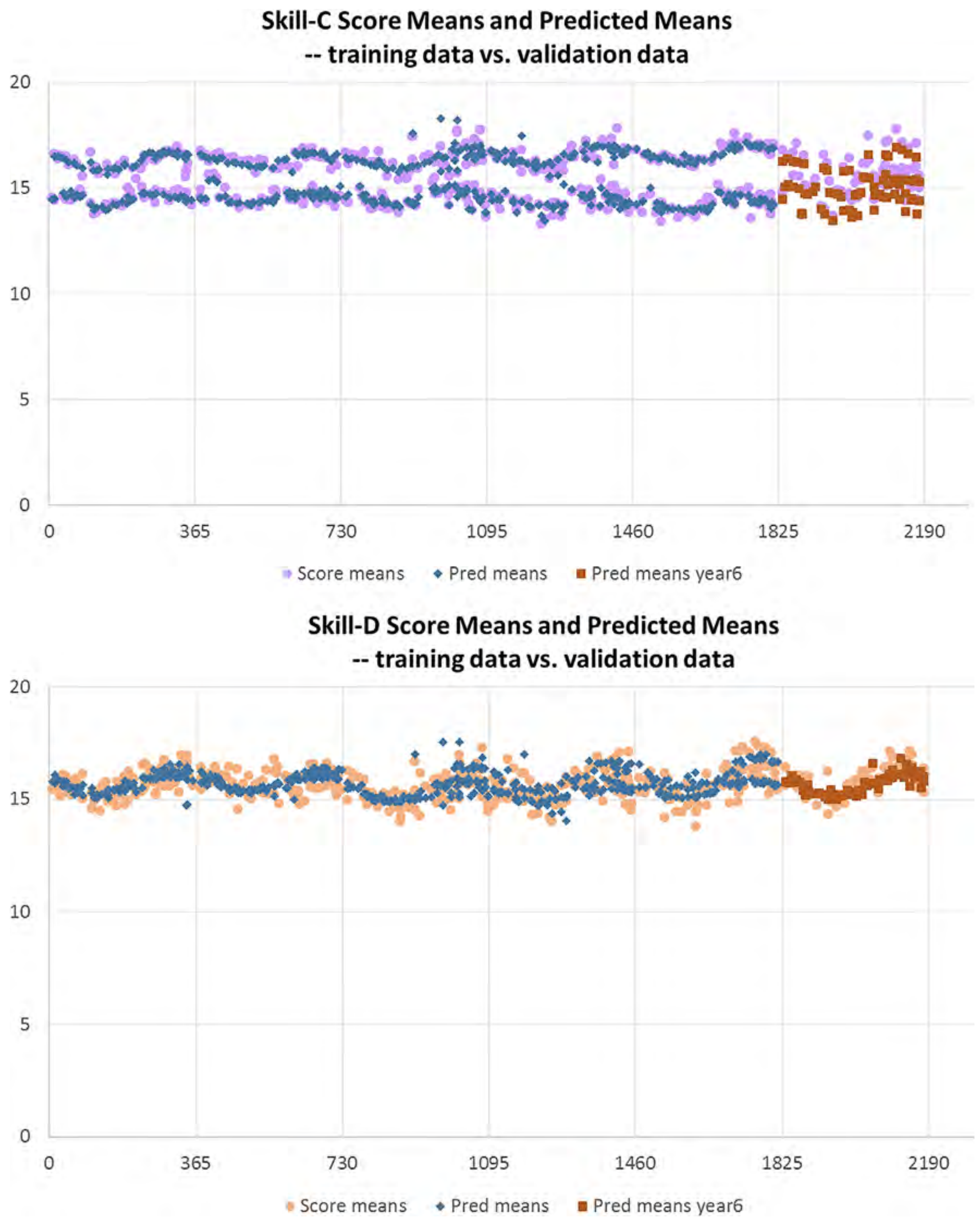

Figure 8 Empirical mean scale scores of each test administration for all 6 years' data for Skills C and D, with dark points representing the predicted means yielded by the training data (Year 1 to Year 5) and the validation data (Year 6); the prediction for Year 6 did not include the predicted year effects in the out-of-sample cross-validation. The test date *t* starts from the beginning time point of Year 1; there are in total 2,191 days (2,191 = 365 × 6 + 1) across Year 1 – Year 6 including 1 day from a leap year (2016).

As mentioned in the Method section, we can apply alternative ways to split the data for the out-of-sample cross-validation—for example, partitioning the data into training data containing Year 1 to Year 4 admins and the validation data containing Year 5 and Year 6 admins. When year is not used as the boundary in partitioning, a number of combinations can be carefully arranged into a grouped jackknifing type design (Haberman et al., 2009; Wang et al., 2018). A rotation design of jackknife grouping, derived from period-based partition, can yield replicated results from each rotation cycle. These replicated results, including year effects in the estimates, can be used to assess the jackknifed prediction accuracy of each point on the cross-validation for sequential data with seasonality. Though details of how to implement the methodology are beyond the scope of the current study, a proposed algorithm for the rotation design is provided in detail for interested readers in Appendix A, An Algorithm of Assessing the Point Prediction Errors in the Out-of-Sample Cross-Validation with Sequential Data. The empirical results and improvement of the algorithm should be a topic for future research.

Model Adequacy Checking: Model Overfitting

Because the R^2 of the HR models for Skill C was high (> 0.9), one concern was whether the high value of the R^2 was due to model overfitting, or, alternatively, whether Skill C data properly confirmed the HR model in Equation 1. As mentioned, an overfitted regression is tailor-made and cannot be generalized to the future data sets. The R-squared value, regression coefficients, and p values of an overfitted model could be misleading because the model would describe the pattern of random errors in the data rather than the genuine relationship between the dependent and independent variables of interest. It is critical to ensure that the high R-squared values of the fitted models are authentic and not due to model overfitting.

In Table 3, three values of R^2, R^2_{adj}, and R^2_{pred} were very close to each other for the models on both the Year 1–Year 6 full data and Year 1–Year 5 training data. The differences between R^2_{adj} and R^2 were small (< 0.025) for the selected models as well as the full models. The same was true for R^2_{pred} and R^2_{adj} (< 0.025). See Figure 9 for the comparisons. In Figure 10, the differences between RMPRESS and RMSE were less than 4% and 5% for the selected models and full models, respectively. The results showed no evidence of model overfitting for Skill C, indicating the alternative—a high level of explanatory and prediction power of the yielded models. This was supported by the plots of the Skill C data in Figure 2, which displayed clear curves across 6 years with no detached outliers. Moreover, the plots of Skill C in Figure 6, based on the out-of-sample cross-validation, validated the HR model in Equation 1; that is, the predicted points on the training data closely conformed to the periodic shape of the observed points, and the predicted points on the validation data extended the trend of the observed points, following the same shape of their periodic variation.

By the AIC criterion, all differences between the full models and the selected models in Figure 11 were less than 1.5%, for all four skills. Moreover, for all four skills, the AICs of the models for the Year 1 to Year 5 data were larger than the same models with the Year 6 admins added.

Table B2 in Appendix B contains major features of the methods employed for a thorough investigation in this section.

Summary

To ensure that HR models are adequately implemented for data exhibiting seasonality and periodicity, it is necessary to perform appropriate model adequacy checking. There is a void in the literature on how to conduct a comprehensive model adequacy evaluation when applying HR models to sequential data with seasonality. The current study contributes to the existing literature by filling in this gap.

A few statistical tools were used for model adequacy checking, such as graphics, out-of-sample cross-validation prediction, and summary statistics. The harmonic term pairing principle was first proved. Based on this principle, a process was proposed to determine the *full* H in HR modeling. A two-stage strategy was used to select parsimonious models and determine the significant predictors. The first stage involved determining the *full* H of an HR model and selecting the region effects through the backward selection procedure. The second stage contained the significance tests of region effects from the shorter variable list obtained from the first stage. The tests with the Bonferroni correction on a shorter variable list were thus less conservative and had a smaller rate of false negatives.

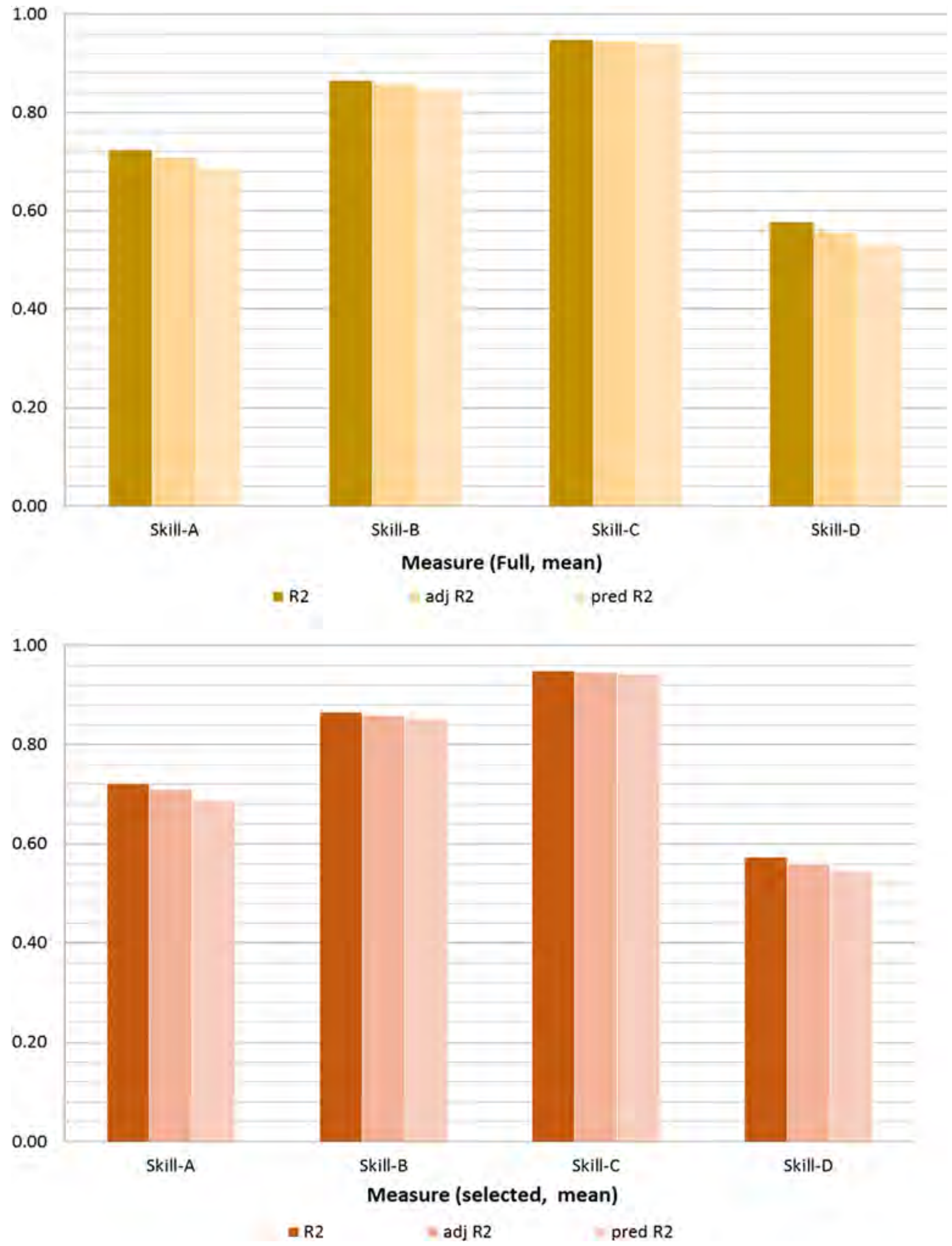

Figure 9 Comparison of three types of *R*-squared of the models for four skills (above, full models; below, selected models).

To effectively implement the out-of-sample cross-validation for sequential data with seasonality, the study used a period-based partition, similar to a grouped jackknifing strategy. Such a strategy is analogous to *PRESS* yet leaves an integrated sub data set out in each rotation cycle. Otherwise, it will be difficult to detect a trend in the predicted points, as they are mixed with observed points that are subject to periodic variation. Moreover, the year effects on the dropped-out points are also included in prediction, as displayed in Figures 5 and 6; as a comparison, those results in Figures 7 and 8 without including the year effects fell farther away from the dots representing the mean score. Alternative ways to split the data for the out-of-sample cross-validation are also suggested. A rotation design of jackknife grouping has been proposed to estimate the jackknifed variance of prediction at each admin point with the point itself not included in modeling for sequential data with seasonality. An algorithm of the rotation design is provided in Appendix A. The statistic *RMPRESS*

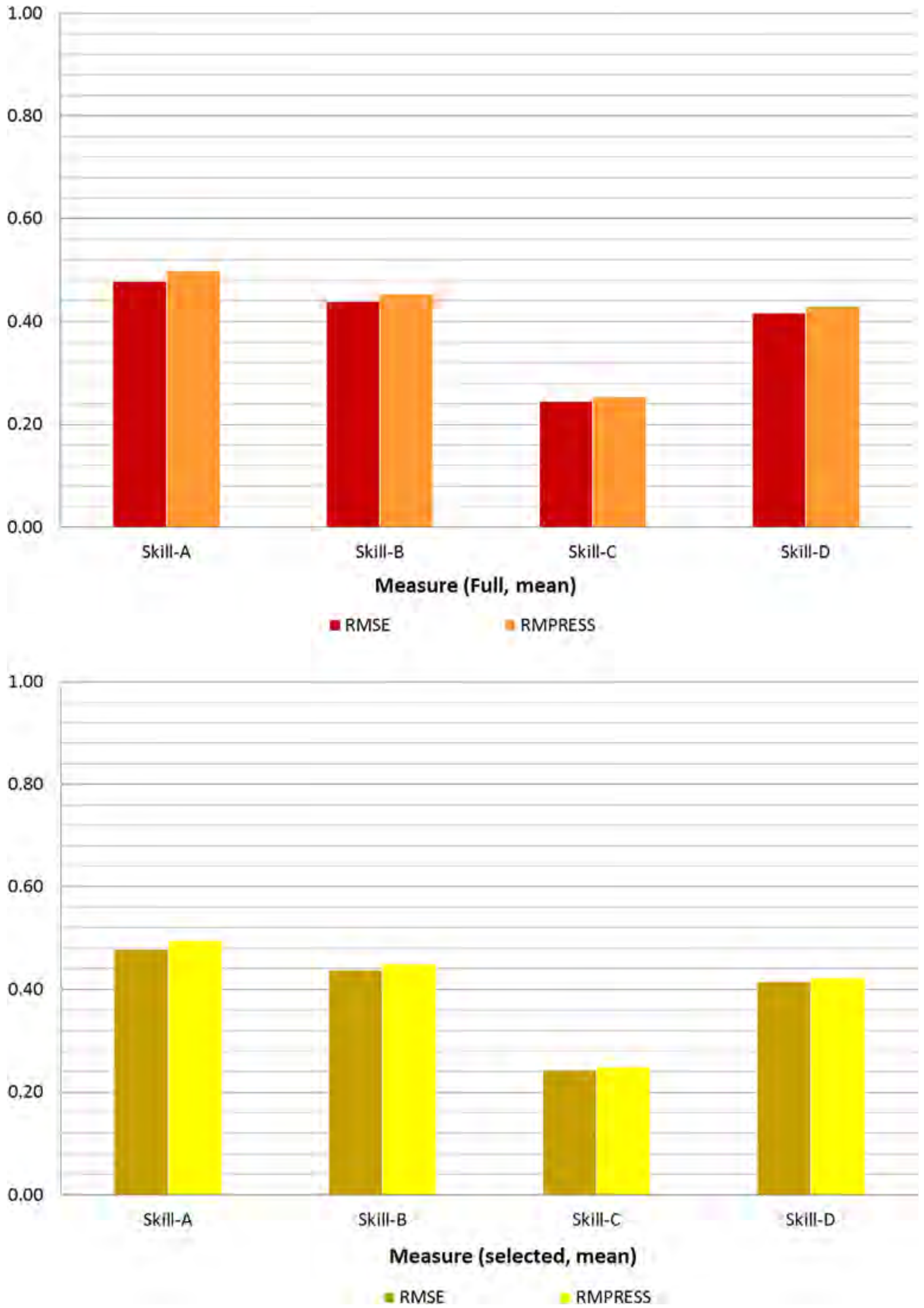

Figure 10 Comparison of root mean squared errors of prediction and root mean predicted residual error sum of squares of the models for four skills (above, full models; below, selected models).

estimates the average prediction error using leave-one-out cross-validation for HR models. It can be directly compared with three types of *R*-squared, RMSE of prediction.

In our analysis, the R^2 of the fitted models for Skill C were very high in value, greater than 0.9. A due diligent check was warranted regarding whether the HR model in Equation 1 can be used to QC the Skill C data, that is, whether the high R^2 was due to model overfitting (Babyak, 2004). The analysis and comparisons of the R^2, R^2_{adj}, and the R^2_{pred} indicated that the model fitting for Skill C was due to remarkably high explanatory and prediction power of the yielded models rather

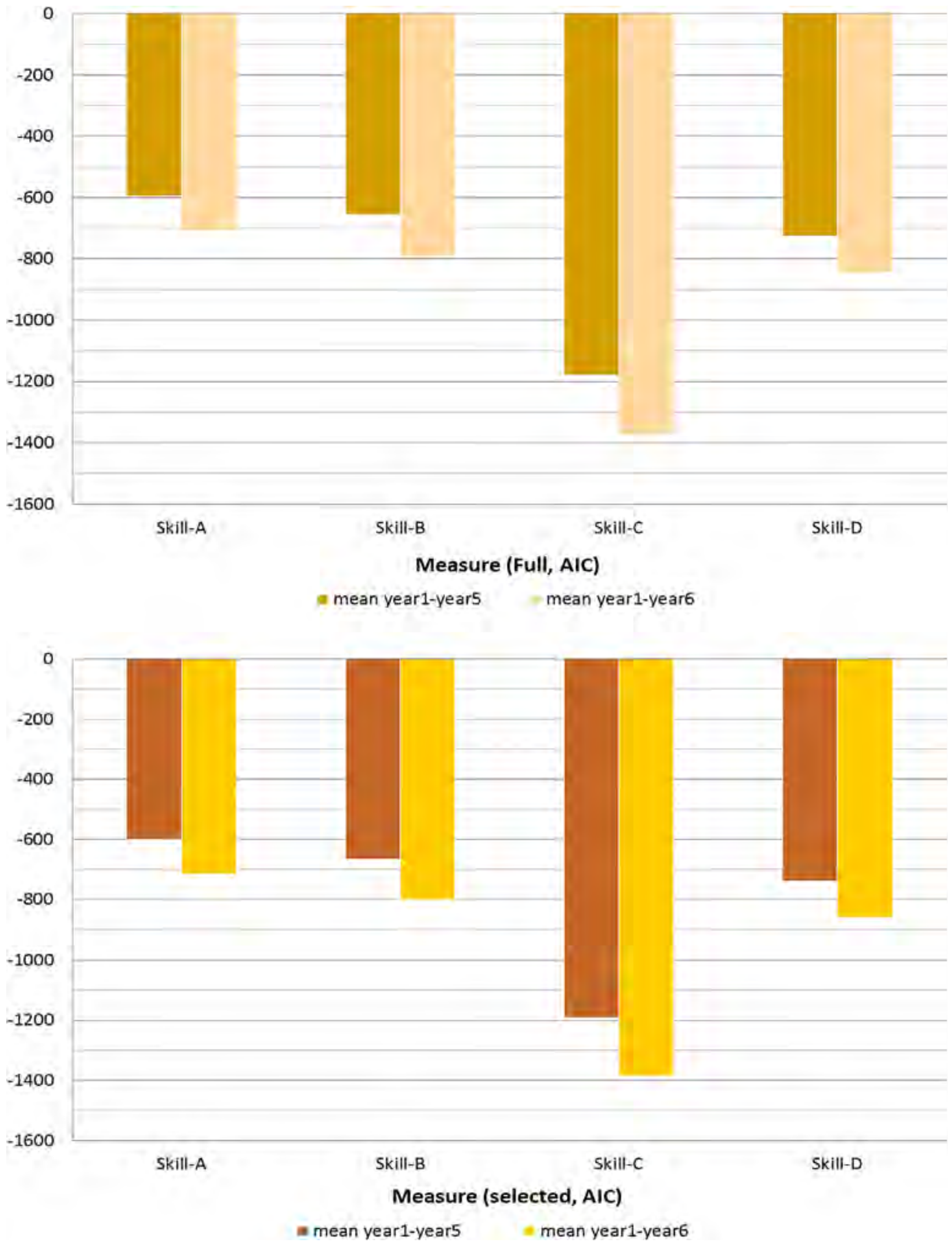

Figure 11 Comparison of Akaike's information criterion for the models for four skills, Year 1 – Year 5 admins versus Year 1 – Year 6 admins (above, full models; below, selected models).

than model overfitting. The result was also confirmed by the cross-validation analysis in the Model Adequacy Checking: Out-of-Sample Cross-Validation section.

In general, the results of the cross-validation analysis demonstrated that the selected HR models were adequate for the data. The model adequacy checking revealed neither signs of poor prediction by the fitted models nor those of model overfitting.

For future studies, to further improve the accuracy of predicted means yielded by the validation data, enhanced forecasting techniques can be used to improve the prediction of the year effects. In addition, we can consider using case weights to obtain a more generalized model that has a weighted distribution that is consistent with the target population distribution of test takers. See Lee and Haberman (2021) for an up-to-date study related to this topic.

Acknowledgments

The authors thank Shelby Haberman, Matthew Johnson, Yi-Hsuan Lee, Hongwen Guo, Nuo Xi, and Sooyeon Kim for their suggestions and comments. The authors also thank Kim Fryer and Ayleen Gontz for their editorial help. Any opinions expressed in this paper are those of the authors and not necessarily those of Educational Testing Service.

Notes

1 The test date t starts from the beginning time point of Year 1; there are in total 2,191 days ($2{,}191 = 365 \times 6 + 1$) across Year 1–Year 6 including 1 day from a leap year (2016).

2 Akaike's information criterion (AIC). For a sample of n administrations, the formula of AIC is $AIC_p = n \ln(SSE) - n \ln(n) + 2p$. In the formulas, n = sample size and p = number of regression coefficients in the model being evaluated (including the intercept). Notice that the Bayesian information criterion is close to the AIC and has a difference in the multiplier of p, which equals $\ln(n)$.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–81). Akademiai Kiado.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, *16*(1), 125–127. https://doi.org/10.1080/00401706.1974.10489157

Andrews, B. J. (2013). *Harmonic regression for TOEFL iBT 2012* [Internal memorandum]. ETS.

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, *5*(2), 141–160. https://doi.org/10.1080/00401706.1963.10490071

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Artis, M. J., Clavel, J. G., Hoffmann, M., & Nachane, D. M. (2007). *Harmonic regression models: A comparative review with applications* (Working Paper No. 333). Institute for Empirical Research in Economics, University of Zurich. https://doi.org/10.2139/ssrn.1017519

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, *89*(425), 177–189. https://doi.org/10.1080/01621459.1994.10476459

Cook, R. D., & Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association*, *92*(438), 490–499. https://doi.org/10.2307/2965698

Courant, R. (1937). *Differential and integral calculus* (2nd ed., Vol. 1). (E. J. McShane, Trans.). New York: Interscience.

Cox, D. R. (2002). Karl Pearson and the chi-squared test. In C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, & M. Mesbah (Eds.), *Goodness-of-fit tests and model validity* (pp. 3–8). Birkhauser. https://doi.org/10.1007/978-1-4612-0103-8_1

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*(3), 161–182. https://doi.org/10.1037/h0025471

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118625590

Durbin, J., & Watson, G. S. (1971). Testing for serial correlation in least squares regression. *Biometrika*, *58*(1), 1–19. https://doi.org/10.2307/2334313

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560. https://doi.org/10.2307/2965703

Gaffney, M., Taylor, C., & Cusenza, E. (1993). Harmonic regression analysis of the effect of drug treatment on the diurnal rhythm of blood pressure and angina. *Statistics in Medicine*, *12*(2), 129–142. https://doi.org/10.1002/sim.4780120205

Guo, H., Robin, F., & Dorans, N. (2017). Detecting item drift in large-scale testing. *Journal of Educational Measurement*, *54*(3), 265–284. https://doi.org/10.1111/jedm.12144

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). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02153.x

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). ETS. https://doi.org/10.1002/j.2333-8504.2009.tb02196.x

Harrell, F. E., Jr. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer. https://doi.org/10.1007/978-1-4757-3462-1

Hillmer, S. C., & Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association*, *77*(377), 63–77. https://doi.org/10.1080/01621459.1982.10477767

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*(1), 1–49. https://doi.org/10.2307/2529336

Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T. (1980). *The theory and practice of econometrics* (1st ed.). John Wiley and Sons.

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*(4), 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & Haberman, S. J. (2021). Studying score stability with a harmonic regression family: A comparison of three approaches to adjustment of examinee-specific demographic data. *Journal of Educational Measurement*, *58*(1), 54–82. https://doi.org/10.1111/jedm.12266

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*, pp. 557–575. https://doi.org/10.1007/s11336-013-9317-5

Li, S., & Qian, J. (2018). *Harmonic regression: Analyzing the regional effects on English-language assessment—An investigation of TOEFL iBT from 2012 to* 2017 [Unpublished manuscript]. ETS.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119013563

McClellan, M., & Staiger, D. (2000). Comparing the quality of health care providers. In Alan M. Garber (Ed.), *Frontiers in health policy research* (Vol. 1, pp. 113–136). MIT Press. https://doi.org/10.1162/109623100300091096,

Pierce, D., & Schafer, D. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, *81*(396), 977–986. https://doi.org/10.1080/01621459.1986.10478361

Qian, J., von Davier, A., & Jiang, Y. (2013). Achieving a stable scale for an assessment with multiple forms: Weighting test samples in IRT linkings. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Springer proceedings in mathematics & statistics* (pp. 171–185). Springer. https://doi.org/10.1007/978-1-4614-9348-8_11

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B*, *31*(2), 350–371. https://doi.org/10.1111/j.2517-6161.1969.tb00796.x

SAS Institute. (2008). SAS/STAT 9.2 user's guide: The REG procedure (Book excerpt). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.3282&rep=rep1&type=pdf

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics*, *19*(4), 415–428. https://doi.org/10.1080/00401706.1977.10489581

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2): 111–147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician*, *54*(2), 116–118. https://doi.org/10.1080/00031305.2000.10474522

Tolstov, G. P. (1976). *Fourier series*. Courier-Dover Publications.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Wang, L., Qian, J., & Lee, Y-H. (2018). *Grouping effects on jackknifed variance estimates for IRT scaling and equating* (Research Report RR-18-16). ETS. https://doi.org/10.1002/ets2.12204

Weisberg, S. (2014). *Applied linear regression* (4th ed.). John Wiley & Sons.

Wolter, K. M. (2007). *Introduction to variance estimation*. Springer.

Appendix A

An Algorithm of Assessing the Point Prediction Errors in the Out-of-Sample Cross-Validation With Sequential Data

In this study, a rotation design of jackknife grouping is proposed under the framework of period-based partition for the out-of-sample cross-validation, and an algorithm has been developed to assess the jackknifed prediction errors at each point for sequential data with seasonality.

Let $\mathbb{R}$ be the sequential data set used in analysis with K indexed cases. In assessment QC, the cases in a dataset can be admins as in the samples used in this paper. The total cases in $\mathbb{R}$ confirm a sequential order $k = 1, 2, \dots K$. To simplify the notations in the algorithm, assume the yearly cases for assessment are the same and the factor of a leap year is not considered. Let M be the number of cases in the sequence dropped in the out-of-sample cross-validation; usually, M can be the number of yearly admins.

Let $v = \text{Mod}\,(b + M - 1, M)$ be a modulo function, where $b + M - 1$ is *number*, M is *divisor*, and $b = 1, 2, \dots K$. Create a sub data set $\mathbb{R}_{(b)}$ $(b = 1, 2, \dots K)$ from $\mathbb{R}$ with a sequence of cases, indexed from b to v, being dropped. That is, M

cases are dropped. Because v is employed in creating $\mathbb{R}_{(b)}$, if the sequential index is larger than K, the sequential index will be equal to the remainder, that is, rotating back to the beginning of the data set.

For each round of HR model fitting under the rotation design, the first step is to create the jackknife replicated results from each rotation round. Because of the modulo function, each case will have M replicated values for sequential data with seasonality. This is the feature of a rotation design, though the algorithm is analogous to repeated resampling approach (Wolter, 2007). The second step is to estimate the jackknifed variances; the estimated variances include the errors due to different factors in the estimation, such as year effects etc.

The rotation of fitting HR regression is based on data $\mathbb{R}_{(b)}$ $(b = 1, 2, \ldots K)$. The total number in the rotated HR model fitting is K times. Therefore, the algorithm described below is computation intensive.

The proposed algorithm is composed of the following steps.

a. For $b = 1, 2, \ldots K$, we complete the following process. The first round $(b = 1)$ of HR fitting is based on data $\mathbb{R}_{(1)}$. Define batch number $u = \text{INT}\left[(k - b)/M\right]$ where INT $[\cdot]$ is integer function and $k = 1, 2, \ldots K$. In the first round, the batch numbers, INT $\left[(k - 1)/M\right]$, are assigned to each case. For all the cases, we yield a set of predicted values. Let $\widehat{y}_{(k,1,u)}$ be the predicted value yielded in Round 1 for the admin with index k and batch number u. The vector $(k, 1, u)$ is called the status segment of $\widehat{y}_{(k,1,u)}$. The predicted value set is $\left\{\widehat{y}_{(k,1,u)} \mid k = 1, 2, \ldots K\right\}$. Then start the second round $(b = 2)$ of fitting HR model and yielding the batch number u for each case and the set of the predicted values $\left\{\widehat{y}_{(k,2,u)} \mid k = 1, 2, \ldots K\right\}$. Continue this process until $b = K$.

b. For admin point $k = k_0$, based on the status segments of all the predicted values for the admin with index k_0, define $R_{k_0,b,0} = \left\{\left(k_0, b, u\right) \mid u = 0, \text{for any } b\right\}$; the size of $R_{k_0,b,0}$ is M.

c. For $k = k_0$, calculate the mean of the predicted values in $R_{k_0,b,0}$,

$$\hat{\widehat{y}}_{(k_0,\cdot,0)} = M^{-1} \sum_{(k_0,b,u)\in R_{k_0,b,0}} \widehat{y}_{(k_0,b,u)};$$

and the jackknifed variance of $\widehat{y}_{(k_0,\cdot,\cdot)}$ is estimated by

$$v_{J1}\left(\widehat{y}_{k_0}\right) = \frac{M-1}{M} \sum_{(k_0,b,u)\in R_{k_0,b,0}} \left(\widehat{y}_{(k_0,b,u)} - \hat{\widehat{y}}_{(k_0,\cdot,0)}\right)^2$$

(Haberman et al., 2009; Wolter, 2007). $v_{J1}\left(\widehat{y}_{k_0}\right)$ measures a variance excluding admin point k_0 itself in modeling. The jackknifed standard error is then computed by

$$se_{J1}\left(\widehat{y}_{k_0}\right) = \sqrt{v_{J1}\left(\widehat{y}_{k_0}\right)}.$$

d. Let y_{k_0} be the mean score of admin point k_0. An alternative jackknifed variance estimation is

$$v_{J2}\left(\widehat{y}_{k_0}\right) = \frac{M-1}{M} \sum_{(k_0,b,u)\in R_{k_0,b,0}} \left(\widehat{y}_{(k_0,b,u)} - y_{k_0}\right)^2$$

(Wolter, 2007). $v_{J2}\left(\widehat{y}_{k_0}\right)$ is also derived based on resampling method and measures the overall errors excluding admin point k_0 itself in modeling.

Appendix B

Summary of the Statistics, Processes, and Methods in HR Model Adequacy Checking

Table B1 Statistics Used to Examine Harmonic Regression Model Fit

Statistics	Symbol	Equation in text	Function	Note
Root mean squared error of prediction	RMSE	5	Measures the error of prediction	
R-squared	R^2	6	Examines model fit, measuring how close the data are to the fitted regression line	Conventional way to check goodness of fit
Adjusted R-squared	R^2_{adj}	7	Examines model fit with adjusted degrees of freedom to the squared terms in R^2	To apply when a model has many independent variables
Predicted R-squared	R^2_{pred}	9	Examines model fit based on leave-one-out cross-validation of the terms in R^2	To apply in judging when a model is possibly overfitting
Predicted residual error sum of squares	*PRESS*	8	Measures the predicted residual error sum of squares residual based on leave-one-out cross-validation	Provides an appropriate summary measure of the model fit for prediction
The root mean-*PRESS*	*RMPRESS*	10	Defined as the root of the mean *PRESS*	Same as *PRESS*

Table B2 Processes and Methods Used in Harmonic Regression Model Adequacy Checking and Cross-Validation

Task	Method	Goal	Note
Check sequential data with seasonality	Plot, e.g. Figures 1 and 2	To confirm seasonality in sequential data	Assuring appropriateness of using HR. See sections: Model Selection and Results
Determining the harmonic terms of a HR	The process based on the principle of employing harmonic terms in pairs	To determine the *full H* of a HR	Including validated harmonic terms in HR. Unique for data with seasonality. See sections: Process to Determine the Harmonic Term Pairs and Model Selection
Model adequacy: regular checking	Check R^2 and compare it with R^2_{adj}	To examine model fit, measuring how close the data are to the fitted regression line	Using R^2_{adj} is more appropriate for a model with many independent variables. See sections: Harmonic Regression Model & Model Selection
Model adequacy: out-of-sample cross-validation	Use a subsequence of admins (excluded from training data) to perform cross-validation by comparing its plot with the plot of the predicted values yielded from training data	To evaluate model adequacy based on the out-of-sample data. In the example, the training data consist of Year 1 to Year 5 and the validation data consist of Year 6	Conducting cross-validation by comparing the plots based on the training and out of training samples, focusing on prediction. Unique for sequential data with seasonality. See section: Process to Conduct Out-of-Sample Cross-Validation
Model overfitting: leave-one-out cross-validation	Check R^2_{pred} and compare it with R^2_{adj} and *RMPRESS*; check the plots of the mean scale scores for Skill C data in Figure 2 display clear curves across years	To evaluate model adequacy and overfitting by using R^2_{pred} and *RMPRESS*, estimated by leave-one-out data, can be compared directly	Using statistics based on leave-one-out principle is more appropriate in cross-validation for model adequacy, judging model overfitting. See sections: Three Types of R-squared, RMSE of Prediction, and RMPRESS and Model Adequacy Checking: Model Overfitting

Table B3 Processes and Methods Used in Checking Major Regression Assumptions

Task	Method	Goal	Note
The residuals have zero means and are uncorrelated	Evaluate if the residuals have zero means across the predicted values and the random dispersion around the X axis without trends. Also use the residual plot in Figure 3	To confirm the residuals have zero means and are randomly distributed without a trend	See Draper & Smith, 1998; Weisberg, 2014
The residuals are normally distributed	Examine the plot of residuals vs. the predicted values and/or normal QQ plots for the residuals of the HR models in Figure 3. Sometimes use a normal Predicted Probability (P – P) plot	To confirm if the residuals follow a normal distribution	See Draper & Smith, 1998; Weisberg, 2014
Homogeneity of variances	Conduct ANOVA, or examine the scatterplots of the predicted values and residuals, such as in Figures 3 and 4	To confirm if the residuals have constant variances	See Draper & Smith, 1998; Weisberg, 2014
The predictor variables in the regression have a linear relationship with the outcome variable	Check the residual plot without linear or seasonal trends; evaluate normality and homoscedasticicity in residuals; examine the normal Predicted Probability (P – P) plot	To confirm the linearity of regression	See Draper & Smith, 1998; Weisberg, 2014
No apparent autocorrelation between adjacent cases in the residuals (Draper & Smith, 1998)	Statistical test based on Durbin – Watson *d* statistic	To detect autocorrelation, correlation	See Draper & Smith, 1998; Durbin & Watson, 1971; Weisberg, 2014

Suggested citation:

Qian, J., & Li, S. (2021). *Model adequacy checking for applying harmonic regression to assessment quality control* (Research Report No. RR-21-13). ETS. https://doi.org/10.1002/ets2.12327

Action Editor: Sooyeon Kim

Reviewers: Hongwen Guo, Yi-Hsuan Lee, and Nuo Xi

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/