JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
DOI: 10.1111/jedm.12227
Journal of Educational Measurement
Fall 2019, Vol. 56, No. 3, pp. 610–625

Predicting Operational Rater-Type Classifications Using Rasch Measurement Theory and Random Forests: A Music Performance Assessment Perspective

Brian C. Wesolowski
University of Georgia

*The purpose of this study was to build a Random Forest supervised machine learning model in order to predict musical rater-type classifications based upon a Rasch analysis of raters' differential severity/leniency related to item use. Raw scores (*N = *1,704) from 142 raters across nine high school solo and ensemble festivals (grades 9–12) were collected using a 29-item Likert-type rating scale embedded within five domains (tone/intonation,* n = *6; balance,* n = *5; interpretation,* n = *6; rhythm,* n = *6; and technical accuracy,* n = *6). Data were analyzed using a Many Facets Rasch Partial Credit Model. An* a priori k*-means cluster analysis of 29 differential rater functioning indices produced a discrete feature vector that classified raters into one of three distinct rater-types: (a) syntactical rater-type, (b) expressive rater-type, or (c) mental representation rater-type. Results of the initial Random Forest model resulted in an out-of-bag error rate of 5.05%, indicating that approximately 95% of the raters were correctly classified. After tuning a set of three hyperparameters (*n_{tree}, m_{try}*, and node size), the optimized model demonstrated an improved out-of-bag error rate of 2.02%. Implications for improvements in assessment, research, and rater training in the field of music education are discussed.*

In the field of music education, solo and ensemble festivals are an important third-party mechanism for assessing students' abilities to perform musical literature on an instrument. Student participation in solo and ensemble festivals is an important part of the school music experience, and is linked to improved student motivation, higher student self-efficacy, and increased musicianship skills (Austin, 1988; Banister, 1992; Franklin, 1979; Hurst, 1994; Sweeney, 1998). Often, students' yearly musical repertoire is selected specifically for the purpose of participating in these festivals (Crochet, 2006) and results often influence teachers' curricular decision making, educational goals, and performance standards in the music classroom (Abeles, Hoffer, & Klottman, 1994; Howard, 2002).

Although solo and ensemble festivals are not high-stakes in the sense that they influence educational decisions such as grade advancement, graduation, or funding considerations, they do carry considerable weight with students, parents, teachers, and administrators (Hash, 2013). In particular, the results are often a barometer for the perceived success and quality of a music program as well as a key indicator for stakeholder perceptions of music teacher effectiveness (Boyle, 1992; Burnsed, Hinkle, & King, 1985; Kirchhoff, 1988; Sivill, 2004). With the increased attention on evaluating teacher effectiveness of "non-tested" subjects such as music (Buckley & Marion, 2011), suggestions have been made to use results of formal music

performance assessments as part of music teacher effectiveness evaluations (Hash, 2013). Therefore, the validity, reliability, and fairness of formal music performance assessment scoring outcomes are becoming increasingly important in the field of music education.

Best practice in the selection and deployment of adjudicators (i.e., raters) for formal music performance assessments suggests that music content experts offer the best chance for providing a fair and equitable assessment. Criteria to become an active rater often include a certain minimum years of teaching, a past history of consistent student and program success at formal music performance evaluations, and peer nomination based upon perceptions of program status and successful teaching, for example (Florida Bandmasters Association, 2017). Research investigations into music content experts' scoring results, however, suggest that there is considerable variability that can affect the validity, reliability, and fairness of scoring outcomes, regardless of music content expertise and fulfillment of these criteria. Specifically in the context of music performance assessments, evidence of variability has manifested through rater errors (Wesolowski, Wind, & Engelhard, 2016a), raters' use of rating scale categories (Wesolowski, Wind, & Engelhard, 2016b), differential rater functioning (DRF; Wesolowski, Wind, & Engelhard, 2015), DRF over time (Wesolowski, Wind, & Engelhard, 2017), rater accuracy (Wesolowski & Wind, 2019), and differential rater accuracy over time (Wind & Wesolowski, 2018), for example. The results of such investigations not only suggest a clear need for improved psychometric considerations of rater behavior in the field of music, but indicate that the psychometric data can potentially be used to inform and improve rater training protocols.

For many state music education associations, rater training usually lasts for 1 day per year and is workshop-based. As examples, the Florida Bandmasters Association provides multiple trainings in 1 day based on the topics of standards and ethics, solo and ensemble adjudication, concert band adjudication, sight-reading adjudication, and jazz band adjudication (Florida Bandmasters Association, 2018). The New York State School Music Association holds a rater orientation workshop at the start of each school year (New York State School Music Association, 2018a) and uses music teacher feedback forms to monitor rater processes (New York State School Music Association, 2018b). The Wisconsin Music Educators Association requires adjudicators to participate in one workshop once every 4 years to maintain an active rater status (Wisconsin School Music Association, 2018). Unlike rater training and monitoring protocols for high-stakes performance assessments that are highly funded and where rater behavior is consistently monitored (e.g., writing assessment), rater training in music is far less supported. Therefore, it is currently more advantageous to control for each raters' level of severity/leniency in the process of estimating measures of students' musical performance ability. This process, however, assumes that raters' severity/leniency is invariant across items and/or subgroups. In the case that raters' severity/leniency is not invariant across items and/or subgroups, evidence of DRF can go undetected and potentially confound the intended interpretation of the resulting scores (Engelhard, 2008). This is particularly problematic in the case of music performance assessments, where the students' performances and the raters'

scoring occur simultaneously in a live setting, with little to no monitoring of rater training that addresses these concerns of fairness.

A recent study examined music rater-types based upon differential severity and leniency associated with rating scale items, rating scale category functioning, and domains of music performance assessment (Wesolowski, 2017). Using a Many Facets Rasch Partial Credit Model to estimate measures and conduct a rater-by-item DRF analysis, a k-means cluster analysis based on rater-by-item bias indices suggested that there are three distinct music rater-types: (a) the syntactical rater; (b) the expressive rater; and (c) the mental representation rater. The domain-level characteristics for the syntactical rater-type included systematically severe scoring of tone/intonation, balance, interpretation, and technical accuracy. The domain-level characteristics for the expressive rater-type included systematically severe scoring for tone/intonation and interpretation but systematically lenient scoring for rhythm. The domain-level characteristics for the mental representation rater-type included systematically severe scoring of intonation and technical accuracy and systematically lenient scoring for tone, balance, and interpretation. The clustering identified in this study will form the foundation for the rater classification structure used in the supervised machine learning approach to rater classification in this study.

Although k-means cluster analyses were helpful in better understanding tendencies of raters' differential severity/leniency, it was conducted as an *a posteriori* analysis that alone is not helpful in directly improving rater behavior, particularly when rater training occurs so infrequently and empirical investigations of rater effects is not common in the field of music. As rater training is currently conducted in the field of music education, it is advantageous to diagnose potential DRF prior to a rater-training workshop and to address evidence of DRF in breakout sessions with the raters themselves during the workshop. Because of the amount of raters that attend these workshops, the time restraints of the workshops themselves, and the lack of individuals in the field with psychometric expertise, it is virtually impossible to provide enough individualized attention to calibrate, train, and recalibrate raters in hopes of removing potential DRF from the scoring process. One solution is to provide an online mechanism for raters to score exemplar performances with the intent of immediately predicting their rater classification and to address possible DRF with individually classified groups of rater-types in the same workshop. In order to accomplish this, a predictive classification model can be used that can provide immediate diagnostic feedback of rater classifications to stakeholders and facilitators of these rater-training workshops. The purpose of this study was to build Random Forest supervised machine learning model in order to predict rater-type classifications based upon a Rasch analysis of raters' differential severity/leniency related to item use. The research questions that guided this study include the following:

1. How accurately can a Random Forest model classify raters into each of the three rater-type classes?
2. What adjustments to the hyperparameters of the model can be made to maximize the model's error rate?
3. What are the most important items for predicting the classification of raters?

Machine Learning and Random Forests

Machine learning is a branch of artificial intelligence that uses automated, algorithmic systems to yield valuable insights and derive meaning from complex data structures (Mitchell, 1997). Recently, with greater public access to big data, rapid advancements in computational performance, and the ability to glean volumes of data so massive that it surpasses the ability of humans to make sense of it, machine learning systems have provided a fruitful method for automating data-driven environments that learn from data through pattern recognition and, more importantly, use pattern recognition to learn from changes in data. Popular applications of machine learning include voice recognition systems, facial detection systems, search engine queries, rideshare costs, email spam filters, and movie recommendation systems (Mohammed, Khan, & Bashier, 2017).

Supervised machine learning, a branch of machine learning, refers to a family of algorithms that train a statistical model of known input and output data with the intent of predicting uncertain, future outputs (Kotsiantis, 2007). Supervised machine learning algorithms analyze a certain percentage of a data set (i.e., training data) to produce an inferred function that can be used to map predicted labels (in the case of categorical outputs for classification problems) or predicted scores (in the case of continuous outputs for regression problems to new input data. The benefit of building and tuning a supervised machine learning model is the ability to provide new sets of categorical or continuous predictions using new and unfamiliar sets of input data without changes to the model specifications. One supervised machine learning algorithm is Random Forests (Breiman, 2001). Random Forests is considered to be one of the most popular supervised learning algorithms because of its high predictive accuracy, ease of interpretability, and ability to detect variable importance measures.

In the case of classification problems, specifically, the Random Forests algorithm is built from the foundational premise of using a single, binary classification referred to as a decision tree, to make predictive classification decisions. A decision tree is a predictive input-output model represented by a tree structure that takes its values from a series of individual input variables that contain a sample of observations of similar response values (acting as predictors) and an individual output variable (acting as an a priori classifier). The parent nodes of the tree represent a subset of the predictor space that are then partitioned into various subsets of children nodes. The partitioning of the nodes is done through a recursive feature whereby the parent nodes and their respective children nodes maximize the decrease in impurity (i.e., a measure on which the optimal decision is based). Terminal nodes, or leaves, are labeled by an empirically calculated best guess of the classified outcome variable.

For data sets with large numbers of input variables, implementation of decision trees as a single support tool for decision making has many drawbacks for model implementation, including high variance, overfitting of the model, and difficult interpretability. The Random Forests algorithm, however, is an ensemble method that combines several base, decision-tree models, or forests, that are built from a series of recursive partitioning of the data that can overcome many of the limitations of a single binary decision tree. Random Forests is an ensemble method that relies on multiple decision trees to produce an optimized predictive model. Random Forests

was selected as an appropriate supervised machine learning model in this study for several reasons: (a) its ability to predict a multi-class outcome variable, (b) its ease of interpretability, (c) its ability to calculate measures of variable importance, and (d) its safeguard against model-to-data overfit as decision trees are added to the forest.

As outlined by Liaw and Wiener (2002), the Random Forest classification works conceptually as follows:

1. Draw n_{tree} bootstrap samples from the original data (where n_{tree} represents the optimized number of trees in the forest to grow).
2. For each of the bootstrap samples, grow an unpruned (i.e., a nonlimiting number of decision trees) classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables (where m_{try} represents the optimized number of random variables initially sampled at each split).
3. Predict new data by aggregating the predictions of n trees (i.e., majority votes) (p. 18).

The feature vectors are trained, aggregated, and classified on randomized bootstrapped samples of the training data. Each tree within a forest provides an estimated classification based upon a random subset of features. The prediction function for a forest is defined as the sum of the feature contributions plus the mean of the highest tree region that covers the training data set (i.e., "bias").

The technique of sampling a randomized set of input variables is referred to as the random subspace method, or "feature bagging." The use of random subspace methods in supervised machine learning processes allows for reduced correlation between the sampled trees, resulting in better overall model performance. The overall resulting classification of each tree is indicated by a pooled, majority vote of classification predictions. Upon predicting classifications of the training data, an estimate of the error rate, or "out-of-bag error," is obtained by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample using the tree grown with the bootstrap sample; then
2. Aggregate the predictions (Liaw and Weiner, 2002, p. 18).

A key feature of the Random Forest model is its ability to detect a hierarchy of variable importance for making classification decisions. One important way of detecting variable importance is through the Gini variable importance measure (GVIM), or more specifically, the mean decrease Gini measure (Breiman, 2001). These estimations of total decrease in node impurity are carried out tree by tree throughout the algorithm's recursive procedure. Detailed model and impurity specifications can be found in Breiman (2001).

Method

Participants

Raw scores ($N = 1{,}704$) from 142 raters across nine high school (Grades 9–12) solo and ensemble festivals were collected. The raters were all music context experts with an active adjudicator status in the representative state of the solo and ensemble festival. Each rater's specialized instructional area was instrumental music education. Raters' raw scores were kept confidential in order to be used as part of this study. Due to the limitation of using authentic music performance assessments where students were evaluated live by raters, an unbalanced incomplete rater assessment network was used for this study (Engelhard, 1997). At minimum, at least two raters evaluated each performance and all sequentially assigned raters were connected across at least one musical performance.

Apparatus

The measurement instrument used in this study was a variation of DCamp's (1980) rating scale suited specifically for a solo and ensemble assessment context. The rating scale consisted of 29 items embedded within five performance domains (tone/intonation, $n = 6$; balance, $n = 5$; interpretation, $n = 6$; rhythm, $n = 6$; and technical accuracy, $n = 6$). The response categories were based upon a 4 point Likert-type scale response set (*strongly agree, agree, disagree, strongly disagree*).

A Priori Analyses

Building on the work of Wesolowski (2017), the Many Facets Rasch Partial Credit (MFR-PC) model (Linacre, 1989; Masters, 1982) with three facets (persons, raters, items) was used to convert observed, raw scores to linear measures. The PC version of the model was specifically chosen to gather detailed information of the raters' individual use of the category structures across each item. The MFR-PC model in this study was specified as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk}-1}\right] = \theta_n - \lambda_i - \delta_j - \tau_{ik}, \tag{1}$$

where $\ln[P_{nijk}/P_{nijk-1}]$ is the probability that performance n rated by rater i on item j in level m receives a rating in category k rather than category $k-1$; θ_n is the logit-scale location (e.g., achievement) of performance n; λ_i is the logit-scale location (e.g., severity) of rater i; δ_j = the logit-scale location (e.g., difficulty) of item j; and τ_{ik} is the logit-scale location where rating scale categories k and $k-1$ are equally probable for rater i.

In the case that the rater is the object of measurement, there are five requirements for rater-invariant measurement: (a) rater-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular raters that happen to be used for the measuring); (b) noncrossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from raters than a less able person; (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular persons used for calibration);

Table 1
Summary Statistics From the PC-MFR Model

	Facets		
	Performance (θ)	Rater (γ)	Item (δ)
Measure (Logits)			
Mean	−.05	.00	−.05
SD	.62	.39	.42
N	1704	142	29
Infit MSE			
Mean	1.01	0.00	.99
SD	.24	.17	.37
Std. Infit MSE			
Mean	−.10	0.00	−.70
SD	1.90	1.50	3.80
Outfit MSE			
Mean	1.04	1.04	1.04
SD	.28	.24	.47
Std. Outfit MSE			
Mean	.10	.20	−.50
SD	2.00	1.80	3.70
Separation statistics			
Reliability of separation	.98	.95	.98
Chi-square	60167.28*	2487.54*	1212.62*
Degrees of freedom	1703	141	28

*$p < .01$.

(d) noncrossing rater response functions (i.e., any person must have a better chance of obtaining a higher rating from lenient raters than from more severe raters; and (e) variable map (i.e., persons and raters must be simultaneously located on a single underlying latent variable) (Engelhard, 2013). The benefit of using the MFR-PC model is that when adequate fit to the measurement model is actively obtained, rater-mediated invariant measurement is achieved. Summary statistics for the MFR-PC model are found in Table 1, and indicate overall good data fit to the MFR-PC model.

A DRF was conducted in order to examine evidence of raters' differential leniency/severity (Engelhard, 2008). Raters' differential leniency/severity was examined across rating scale items and categories in order to establish rater-by-item bias indices for each item. The interaction term between raters and items ($\lambda_i \delta_j$) was added to the MFR-PC model as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = (\theta_n - \lambda_i - \delta_j - \tau_{ik}) - \lambda_i \delta_j, \quad (2)$$

where $\lambda_i \delta_j$ is the interaction between rater severity and item difficulty.

A total of 4,118 rater-by-item interactions were computed from the DRF analysis. The analysis indicated an overall statistically significant parameter-level differential measure ($\chi^2_{(4118)} = 4921.68, p < .01$), explaining 29.64% of the variance within the

measurable responses. *FACETS* (Linacre, 2014) was used to estimate measures and conduct DRF analyses.

A nonhierarchical *k*-means cluster analysis of the item bias indices for each rater was conducted in a previous study (Wesolowski, 2017) in order to classify each rater into one of three possible classifications: (a) syntactical, (b) expressive, or (c) mental representation. The same cluster seeds were used in order to prespecify threshold distances that demonstrated reproducibility of the initial analysis. Based on the analysis, raters from this study were classified into three possible rater-type classifications: (a) syntactical rater-type ($n = 62$), (b) expressive rater-type ($n = 33$), or (c) mental representation rater-type ($n = 32$). The *k*-means analysis was conducted in R statistics software (R Core Team, 2017) using the base R hclust function and the hsmic package (Harrell, 2017).

Random Forest Data Set Preparation

The results of the rater-by-item DRF analysis and cluster analysis were established as a dataframe in R statistics software consisting of 142 observations (i.e., raters) and 30 variables. In the context of standard machine learning vernacular (Google Developers, 2018), the 30 variables (29 item bias indices and rater classification assignments) are referred to as input variables, otherwise referred to as feature vectors. Grouped together, the 30 feature vectors are referred to as a feature set. The feature set consisted of one discrete feature vector (the classification of the rater based upon the k-means cluster analysis) and 29 continuous feature vectors (the item bias indices). In the case of the Random Forest model built in this study, the 29 continuous feature vectors were used to make predictions based upon the discrete feature vector results (the rater classifications). The feature (i.e., output) column of the dataframe was specified as rater-type, consisting of three possible classifications: (a) syntactical rater-type, (b) expressive rater-type, and (c) mental representation rater-type. The remaining 29 feature vectors were specified as continuous feature vectors (raters' bias indices for each 29 items).

The dataframe was randomly subset into a 70% training data set ($n = 99$ raters) and 30% testing data set ($n = 43$ raters). Three hyperparameters were included as part of the model specifications as suggested by Breiman (2001). First, the initial number of trees to grow (i.e., n_{tree}) specified in the initial model was set to 500 trees, the default setting of the randomForest R package. Second, the number of random variables initially sampled at each split (i.e., m_{try}) was set to five variables, which by default is the rounded square root of the number of input variables included in the data set ($n = 29$). Finally, the minimum node size of each tree (i.e., node size) was set to one node, also the default setting of the randomForest R package. The model was built using a combination of R packages, including randomForest (Breiman, Cutler, Liaw, & Wiener, 2018), Caret (Kuhn, 2018), and Metrics (Hamner, Frasco, & LeDell, 2018).

Results

Default Model Training

The out-of-bag (OOB) estimate of error rate is a useful tool to broadly evaluate the overall model performance. The OOB is the average prediction error for each of the

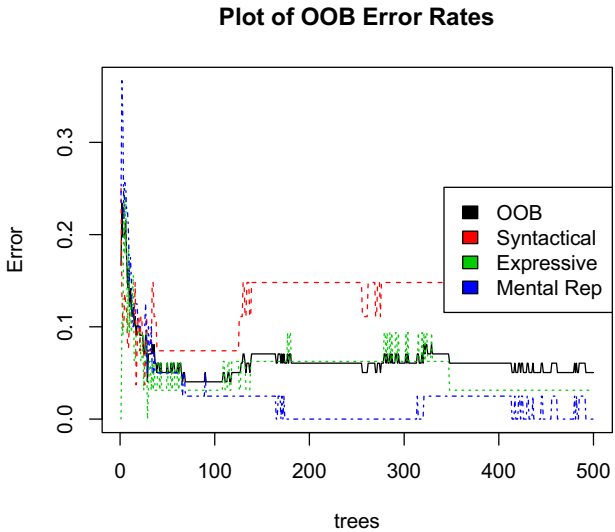

Figure 1. Plot of average out-of-bag error for each rater classification. (Color figure can be viewed at wileyonlinelibrary.com)

bootstrapped samples of the training observations relative to the trees added to the forest. The initial model specifications consisted of three hyperparameters (n_{tree}, m_{try}, and node size) set to the default settings. Based upon the initial model specifications of 500 trees, m_{try} of 5, and node size of 1, the OOB error rate for the initial model was 5.05%. Figure 1 is a plot of the OOB error rates for each rater-type classification as a function of the number of trees specified in the model. Evaluation of the overall OOB error rate provides insight into the optimal amount of trees to be specified that can lead to the model's predictive performance. The optimal amount of trees for this model will be discussed later in the manuscript when the hyperparameters of the initial Random Forest model are evaluated and tuned.

Classification Results and Summary Statistics for the Test Data Set

Model accuracy is defined as the ratio between the number of correct predictions (if yes, then 1; if no, then 0) and the number of observations, or raters, in the data. The overall test accuracy of the model was 88.37% (95% CI [0.75, 0.96]) indicating that only 5 of the 99 cases in the training data set were misclassified. The results of the raw predicted test data classifications can be found in Table 2. The columns of the table refer to the actual classifications and the rows of the table refer to the predicted classifications. The diagonal indicates the true positives and true negatives. The top right of the diagonal indicates the false positive classifications, and the bottom left of the diagonal indicates the false negative classifications.

Summary statistics for each of the rater-types can be found in Table 3. Sensitivity is defined as the proportion of relevant results out of the number of samples that were actually relevant. Specificity is defined as the proportion of true negatives

Table 2
Confusion Table for Initial Model Specification and Tuned Hyperparameter Model

Actual Classifications	Predicted Classifications: Syntactical	Expressive	Mental Representation	Class Error (%)
Initial model specification				
Syntactical	23	2	2	.15
Expressive	1	31	0	.03
Mental representation	0	0	40	<.01
Tuned hyperparameter model				
Syntactical	25	1	1	.07
Expressive	0	32	0	<.01
Mental representation	0	0	40	<.01

Table 3
Summary Statistics for the Initial Random Forest Model by Classification-Type

	Syntactical	Expressive	Mental Representation
Sensitivity	1.00	.81	.88
Specificity	.94	.93	.96
Positive predicted value (PPV)	.82	.87	.94
Negative predicted value (NPV)	1.00	.89	.92
Prevalence	.21	.37	.42
Detection rate	.21	.30	.37
Detection prevalence	.26	.35	.40
Balanced accuracy	.97	.87	.92

that are correctly identified by the test. Positive predictive value (PPV) is the proportion of rater-types matching the corresponding rater-type correctly classified. Negative predictive value (NPV) is the proportion of rater-types not matching the corresponding rater-type correctly classified. Prevalence is interpreted as how often each category occurs in the population. The detection rate is defined as the rate of true events also predicted by the events. Detection prevalence is defined as the commonness of predicted events. Balanced accuracy is defined as the average accuracy obtained for all classes.

Tuning and Selecting Hyperparameters

The default model training was an important step to explore the baseline predictability for rater-type classifications. However, training a sequence of models using a variety of hyperparameter settings and combinations is an important next step for building an optimized model in which OOB error is minimized and model performance is improved. Three important hyperparameters of a Random Forest model

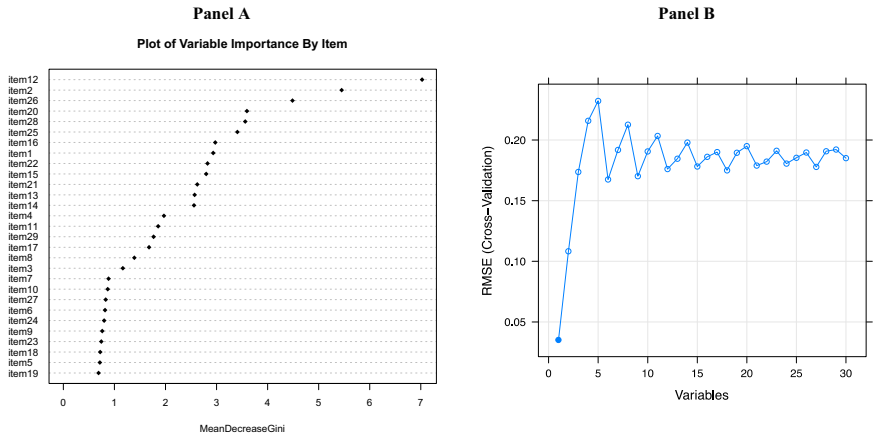

Figure 2. Plot of variable importance by item and the recursive feature vector selection in order of items of importance. (Color figure can be viewed at wileyonlinelibrary.com)

were examined that can have a significant impact on the model performance: (a) the number of trees in the forest (i.e., n_{tree}), (b) the number of variables randomly sampled as candidates for each split (i.e., m_{try}), and (c) the minimum number of samples on the terminal nodes (i.e., node size).

In order to tune the hyperparameters of the model, the tuneRF function with the doBest parameter set true was used from the randomForest R package. Results of the analysis indicated that n_{tree} with the minimum amount of error was 500 trees. Then, a manual grid search was conducted. The process of conducting a manual grid search included establishing a dataframe of all possible combinations of values for m_{try} and node size and training iterative models consisting of each combination in order to identify the optimal set of hyperparameters based on minimal OOB error. The results indicated an optimized model with an m_{try} value of 12 and node size value of 3. The optimized model, using the set of tuned hyperparameter specifications of $n_{\text{tree}} = 120$, $m_{\text{try}} = 12$ and node size $= 3$, demonstrated an OOB error rate of 2.03% (see Table 2).

Relative Importance of Variables

Using a recursive partitioning strategy, decision tree algorithms used for classification problems split the input data into multiple sets of divided regions and assigned labels to the split points, testing the predictability and minimizing the error with each iteration along the way. An artifact of this process is in itself a feature selection process, which highlights the most efficient and relevant features for maximizing correctly predicted partitioning of the data. The Gini variable importance measure (i.e., mean decrease Gini) and root square mean square error (RMSE) are two important metrics that provide empirical evidence of these features' importance. Figure 2 is a plot of variable importance (Panel A) and plot of the recursive feature vector selection (panel B) for each of the 29 continuous feature vectors used in the Random Forest model. Panel A reflects the ordering of the items from most relevant features

to least relevant features that efficiently predict the rater classifications. Panel A answers the interpretative question, "How important is each of the features in correctly predicting the rater classification?" Panel B reflects the aggregate of the features' magnitude of the errors in the prediction of raters. Panel B answers the interpretative question, "What is the average prediction error of all the features?"

Based upon the Gini variable importance measure and recursive feature vector selection operation, five items provide the greatest priority, weight, and accuracy for classifying raters. These items included item 12 (*inner parts are too timid*), item 2 (*basic tuning is not good, plays out of tune*), item 26 (*notes in runs are inaccurate*), item 20 (*dotted eighth-sixteenth pattern is inaccurate*), and item 28 (*awkward and difficult passages are not prepared*). These variables can be used to build a less computationally demanding and more easily interpretable model for future use.

Conclusion and Discussion

The purpose of this study was to build a Random Forest supervised machine learning model in order to predict rater-type classifications based upon a Rasch analysis of raters' differential severity/leniency related to item use. The initial model demonstrated an OOB error rate of 5.05%, with a syntactical rater-type class error of .15, expressive rater-type class error of .03, and a mental representation rater-type of $<.01$. After conducting a hyperparameter tuning process focusing on three distinct hyperparameters ($n_{tree} = 120$, $m_{\text{try}} = 12$, and node size $= 3$), the tuned model demonstrated a an out-of-bag (OOB) error rate of 3.03%, with a syntactical rater-type class error of .13, expressive rater-type class error of .03, and a mental representation rater-type of $<.01$. An analysis of variable importance indicated that a subset of 5 items (item 12, item 2, item 26, item 20, and item 28) provides the greatest weight in classifying raters into the three distinct rater types.

The consideration of the subset of five items from the variables of importance analysis may provide an important starting point for rater training considerations. The domain-level characteristics for the syntactical rater-type included systematically severe scoring of tone/intonation, balance, interpretation, and technical accuracy. Specific items related to these domains are item 12 (*inner parts are too timid*), item 2 (*basic tuning is not good, plays out of tune*), item 26 (*notes in runs are inaccurate*), and item 20 (*dotted eighth-sixteenth pattern is inaccurate*). It is suggested that rater-training protocols implement accuracy models that specifically target syntactical rater-type's systematically severe use of these items. The domain-level characteristics for the expressive rater-type included systematically severe scoring for tone/intonation and interpretation but systematically lenient scoring for rhythm. Specific items related to these domains are item 2 (*basic tuning is not good, plays out of tune*) and item 20 (*dotted eighth-sixteenth pattern is inaccurate*). It is suggested that rater-training protocols implement accuracy models that specifically target expressive rater-type's systematically lenient use of these items. The domain-level characteristics for mental representation rater-type included systematically severe scoring of intonation and technical accuracy and systematically lenient scoring for tone, balance, and interpretation. Specific items related to these domains are item 12 (*inner parts are too timid*), item 2 (*basic tuning is not good, plays out of tune*), item 26

(*notes in runs are inaccurate*), item 20 (*dotted eighth-sixteenth pattern is inaccurate*) and item 28 (*awkward and difficult passages are not prepared*). It is suggested that rater-training protocols implement accuracy models that specifically target mental representation rater-type's systematically severe use of items 2, 26, 20, and 28 and systematically lenient use of item 12.

One unique benefit of both Rasch measurement theory and Random Forests is their ability to detect outliers. In the case of Rasch measurement theory, raters detected as either misfit or those demonstrating differential severity/leniency allow for a more qualitative investigation into unique behaviors that the field can learn from. Additionally, in the case of Random Forests, misclassified raters can also provide a different level of understanding through additional qualitative investigations. In moving forward, and given both the value and importance of formal performance assessments in the field of music education, it is suggested that the field spends considerable more time, effort, and resources in moving towards investigating the effects of rater training on rater behavior and related outcome scores and considering the research other fields have provided in regard to automated scoring systems. Evidence of its potential effectiveness may provide grounds to invest more funding and effort in working with raters in this regard. Not only understanding patterns in rater behavior but an understanding of effects of training on these behaviors can improve not only understanding of rater-mediated assessments, but also listening, cognition, and psychology of music participation.

Because the field of music education values divergence of raters' response in the adjudication process, however, it is important to note that rater's unique perspectives as content experts is still an important component of the evaluation process in music. Therefore, it is suggested that the field move cautiously towards any considerations of using automated systems for scoring, such as the field of automated writing evaluation, for example, for use in the context of formal and/or summative music performance assessments. However, the field of music education research can find great value toward the improvement of technical aspects of music performance assessments should scholars in the field decide to become more aware of rater training research, scoring automation research, and the applications of predictive regression and/or classification modeling such as machine learning drawn from the contexts of educational measurement research. The field would greatly benefit from the interdisciplinary, collaborative research of educational measurement experts in these areas.

As the field of music education, and arts in general, becomes more reliant upon data-driven evidence of student achievement and program effectiveness, it will surely be looking more toward the fields of educational measurement and data science to provide insightful methodologies to both improve the validity, reliability, and fairness of music assessment contexts and as a means to discover new empirical patterns underscoring music teaching and learning. The field is just now beginning to understand the psychometric ramifications of scoring outcomes, and Rasch measurement theory has played an important role in such discoveries. Furthermore, the subjective nature of music, teaching, and learning is complex and multifaceted. Improvements in pattern recognition of these complex processes through the use of machine learning algorithms are helping to improve the understanding of music

specifically in regard to performance assessments and more broadly in the context of teaching, learning, and accountability across the United States.

References

Abeles, H. F., Hoffer, C. R., & Klottman, R. H. (1994). *Foundations of music education* (2nd ed.). New York, NY: Schirmer Books.

Austin, J. R. (1988). The effect of music contest format on self-concept, motivation, achievement, and attitude of elementary band students. *Journal of Research in Music Education*, *36*(2), 95–107.

Banister, S. (1992). Attitudes of high school band directors toward the value of marching band and concert band contests and selected aspects of the overall band program. *Missouri Journal of Research in Music Education*, *29*, 49–57.

Boyle, D. J. (1992). Program evaluation for secondary school music programs. *NASSAP Bulletin*, *76*(544), 63–68.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2018). *RandomForest: Breiman and Cutler's Random Forests for classification and regression*. R package version 4.6-14.

Buckley, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects*. Dover, NH: The National Center for the Improvement of Educational Assessment.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, *21*, 22–29.

Crochet, L. S. (2006). *Repertoire selection practices of band directors as a function of teaching experience, training, instructional level, and degree of success* (Unpublished doctoral dissertation). University of Miami, Coral Gables, FL.

DCamp, C. B. (1980). *An application of the facet-factorial approach to scale construction in the development of a rating scale for band performance*. (Unpublished doctoral dissertation). University of Iowa, Iowa City.

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–33.

Engelhard, G. (2008). Differential rater functioning. *Rasch Measurement Transactions*, *21*(3), 1124.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Florida Bandmasters Association. (2017). *Becoming an FBA adjudicator*. Retrieved from https://fba.flmusiced.org/media/1742/fba-prospective-adjudicator-info.pdf

Florida Bandmasters Association. (2018). *Adjudicator training workshops*. Retrieved from https://fba.flmusiced.org/for-directors/adjudicator-info/workshops/

Franklin, J. O. (1979). *Attitudes of school administrators, band directors, and band students towards selected activities of the public school band program* (Unpublished doctoral dissertation). Northwestern State University of Louisiana, Natchitoches.

Google Developers. (2018). *Machine learning glossary*. Retrieved from https://developers.google.com/machine-learning/glossary/

Hamner, B., Frasco, M., & LeDell, E. (2018). *Metrics: Evaluation metrics for machine learning*. R package version 0.1.4.

Harrell, F. E., Jr. (2017). *Hmisc: Harrell Miscellaneous*. R package version 4.0-3.

Hash, P. M. (2013). Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review*, *114*, 163–169.

Howard, R. L. (2002). *Repertoire selection practices and the development of a core repertoire for the middle school concert band* (Unpublished doctoral dissertation). University of Florida, Gainsville.

Hurst, C. W. (1994). *A nationwide investigation of high school band directors' reasons for participating in music competitions* (Unpublished doctoral dissertation). The University of North Texas, Denton.

Kirchhoff, C. (1988). The school and college band: Wind band pedagogy in the United States. In J. T. Gates (Ed.), *Music education in the United States: Contemporary issues* (pp. 259–276). Tuscaloosa: The University of Alabama Press.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In I. Maglogiannis, K. Karpouzls, B. A. Wallace, & J. Soldatos (Eds.), *Emerging artificial intelligence applications in computer engineering* (pp. 3–24). Clifton, VA: IOS Press.

Kuhn, M. (2018). *Caret: Classification and regression training*. R Package version 6.0-81.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2/3*, 18–22.

Linacre, J. M. (1989). *Many facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2014). *Facets*. Chicago, IL: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2017). *Machine learning: Algorithms and applications*. Boca Raton, FL: CRC Press.

New York State School Music Association. (2018a). *2019 NYSSMA adjudicator orientation seminars*. Retrieved from https://www.nyssma.org/wp-content/uploads/2018/11/AOSApplication19-4.pdf

New York State School Music Association. (2018b). *New York State School Music Association adjudicator feedback form*. Retrieved from http://www.nyssma.org/wpcontent/uploads/2016/01/AdjudicatorFeedbackForm15.pdf

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sivill, J. R. (2004). *Students' and directors' perceptions of high school band competitions* (Unpublished doctoral dissertation). Bowling Green State University, Bowling Green, OH.

Sweeney, C. R. (1998). *A description of student and band director attitudes toward concert band competition* (Unpublished master's thesis). University of Miami, Coral Gables.

Wesolowski, B. C. (2017). Exploring rater cognition: A typology of raters in the context of music performance assessment, *Psychology of Music*, *45*(3), 375–399.

Wesolowski, B. C., & Wind, S. A. (2019). Investigating rater accuracy in the context of secondary-level solo instrumental music performance. *Musicae Scientiae*, *23*(2), 157–176.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147–170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Rater analyses in music performance assessment: Application of the Many Facet Rasch Model. In T. S. Brophy, J. Marlatt, & G. K. Ritcher (Eds.), *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356). Chicago, IL: GIA.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016b). Examining rater precision in music performance assessment: An analysis of rating scale structure using the Multifaceted Rasch Partial Credit Model, *Music Perception*, *33*, 662–678.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *212*, 75–98.

Wind, S. A., & Wesolowski, B. C. (2018). Evaluating differential rater functioning accuracy over time in solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *215*, 33–55.

Wisconsin School Music Association. (2018). *Adjudicator workshops*. Retrieved from https://wsmamusic.org/adjudicator-center/workshop

Author

BRIAN C. WESOLOWSKI is Associate Professor of Music Education at the University of Georgia, Hugh Hodgson School of Music, 250 River Rd., Athens, GA 30602; bwes@uga.edu. His primary research interests include rater behavior, scale development, and policy of educational assessment in music.