Rater Analyses in Music Performance Assessment: Application of the Many Facet Rasch Model

Brian C. Wesolowski, The University of Georgia, USA
Stefanie A. Wind, The University of Alabama, USA
George Engelhard, Jr., The University of Georgia, USA

Abstract

The purpose of this study was to describe and illustrate the use of modern measurement theory for evaluating the psychometric quality of a scale designed to measure jazz big band performance achievement. Many of the measurement models used in music education are based on classical test theory. Advances in item response theory, such as the Rasch measurement model, have not been systematically used to examine assessment practices in music education. The data analyzed in this study are based on 22 items created to measure the quality of jazz big band performance. Experienced judges (N = 23) rated a set of recordings of jazz big band performances by middle school, high school, collegiate, and professional big bands. There were four ratings by two raters for each recorded jazz performance. One common rater additionally evaluated all 23 performances. These data were analyzed with the Many Facet Rasch Rating Scale (MFR-RS) Model. The results of this study suggest that, overall, the rating data demonstrate good fit to the MFR-RS model. Illustrative analyses are used to demonstrate detailed examination of unexpected responses related to particular performances, raters, and items as a method for evaluating the psychometric quality of a rating system for musical performances. Implications for research, theory, and practice of assessments in music education are discussed.

In music, panels of expert adjudicators (i.e., raters) often conduct formal evaluations of musical ensembles (Barnes & McCashlin, 2005; Conrad, 2003; Fautley, 2010). Raters are identified and employed in the assessment process based upon specific characteristics that deem them an "expert," including but not limited to years of experience, success in the field of music education, and their ability to identify, diagnose, and prescribe solutions to common performance problems (Kruth, 1970). Rating scales utilized by the raters often allow for polytomous ratings (e.g., poor, fair, good, excellent and superior) of several items (e.g., phrasing, dynamic changes, nuance, tempi, etc.) categorized within several domains (e.g., sound quality, technical accuracy, musicality, etc.) (NAfME, n.d.). Each item serves as a prompt to elicit ordinal indications of each rater's judgment of the performance. The intent of the quantitative

data is to accurately and fairly evaluate each ensemble, while providing helpful diagnostic feedback to improve teaching and learning for future performances (Fox, 1990).

The problem with rater-mediated assessments, however, is that ratings are often associated with characteristics of raters and not necessarily with the performances themselves (Engelhard, 2002). This threat of construct-irrelevant variability can obscure the latent construct being measured (i.e., performance achievement), the measure itself, and the resulting observed scores (Lane & Stone, 2006). Additionally, the assessment paradigm utilized in large group music performance evaluations calls for major concerns of measurement error due to four specific factors: (a) the ability of the ensemble, (b) the difficulty of the task, (c) variability in rater judgments, and (d) the manner in which the rater applies the measurement instrument. There is currently a lack of a reliable and systematic methodology for analyzing and reporting rating quality of rater-mediated assessments in the field of music education. Therefore, a need exists to expand the understanding of psychometric requirements related to rater-mediated assessments in music performance.

Purpose

The purpose of this article is to demonstrate the application of the Many Facet Rasch (MFR) model (Linacre, 1989/1994) to rater-mediated items and analysis data. More specifically, three areas will be discussed: (a) psychometric considerations for rater-mediated music performance assessments; (b) judging plans and data analysis procedures; and (c) demonstration of MFR model analyses. This study is guided by the following research questions: (a) how do ensembles vary in level of performance? (b) how do the rating scale item vary in difficulty? and (c) how do the raters systematically vary in leniency and severity?

Psychometric Considerations for Rater-Mediated Music Performance Assessments

In music education research, the primary measurement model utilized in scale development using rater-mediated assessment data is based upon Classical Test Theory (CTT). Examples include the measurement of musical affect (Asmus, 1985; Edwards & Edwards, 1971; Jellison, 1985; Sandstrom & Russo, 2013; Shaw & Tomcala, 1976), teaching effectiveness (Bergee, 1992), musical environment (Brand, 1985), musical aptitude (Asmus, 1989), musical achievement (Colwell, 1970) and performance achievement (Abeles, 1973; Bergee, 1995; Nichols, 1991; Russell, 2010; Smith, 2009; Zdzinski & Barnes, 2002). Additionally, research studies analyzing the reliability, validity, and accuracy of performance evaluations, adjudicator ratings, and/or specified assessment tools most often utilize measurement models based upon CTT (Bergee, 2003; Bergee, 2007; Bergee & McWhirter, 2005; Bergee & Platt, 2003; Bergee & Westfall, 2005; Boeckman, 2002; Brakel, 2006; Burnsed, Hinkle, & King, 1985; Ciorba & Smith, 2009; Conrad, 2003; Fiske, 1975; Fiske, 1983; Garman, Boyle, & DeCarbo, 1991; Hash, 2012;

King & Burnsed, 2009; Kinney, 2009; Latimer, Bergee, & Cohen, 2010; Morrison, Price, Geiger, & Cornacchio, 2009; Norris & Borst, 2007; Price & Chang, 2005; Radocy, 1976; Saunders & Holahan, 1997; Silvey, 2009). CTT, or "true-score theory," utilizes raw scores gleaned from a pool of examinees to test their relative success or failure on individual items. Known for its relatively weak theoretical assumptions, the CTT measurement model assumes that the observed scores obtained from a measure are comprised of two parts: (a) true score, and (b) measurement error. Item discrimination, or the ability of an item to discriminate between examinees of varying ability levels, is indicated statistically using Pearson product-moment correlation coefficients. In the case of polytomously scored items (e.g., Likert rating scales), adjusted proportion-correct values (*p*-values) and correlation coefficients are utilized in order to indicate item difficulty and overall ability level of an examinee. The major disadvantage of using CTT as a means for analyzing performance assessment data is the sample and test dependency of estimated person parameters (e.g., true scores) and item parameters (e.g., item discrimination and item difficulty). This limits the ability to develop valid and reliable measures and to make informed inferences related to examinee ability and item difficulty that extend beyond the context of a single assessment situation. Extended applications of CTT include generalizability theory (Brennan, 2001), factor analysis (Harman, 1976), and structural equation modeling (Jorsekog, 2007).

In contrast, the Rasch family of measurement models offers a more grounded theory compared to CTT. Rasch measurement theory (Rasch, 1960/1980) is often preferred in scale development as well as the measurement of latent traits in the behavioral, social, and health sciences (Engelhard, 2013). The major benefit of the Rasch model is that, when adequate fit to the model is observed, invariant measurement is achieved. In the context of assessments, invariant measurement implies that the measurement of persons is not influenced by the particular items that they happen to take, and the measurement of items is not influenced by the particular persons by whom they are measured. Rasch models use probabilistic distributions of responses as a logistic function of person and item parameters in order to define a latent trait. In contrast to CTT where raw scores are directly used in the analyses, Rasch measurement theory converts raw scores to a log-odds scale using a logistic transformation. The transformed test score data can then be conceptualized as a dependent variable with multiple independent variables (i.e., facets) of interest, including measures of rater severity and leniency, item difficulty, task difficulty, and performance achievement level. Hierarchies of difficulty for each relevant item, and each examinee's discrete item responses are mapped onto a single logit (log-odds units) scale. As a result of the mapping of facets onto a single continuous latent variable scale, it is possible to construct a variable map to use as a visual display for illustrating relative differences in locations among facets.

It is important to note that the property of invariant measurement that characterizes the Rasch model must be evaluated in empirical data. Invariant measurement a hypothesis that must be confirmed or disconfirmed by evidence in a data set (Engelhard, 1994). Engelhard and Perkins (2011) provided a set of five

requirements that can be used to determine the degree to which invariant measurement is obtained for persons and items. These requirements include (a) item-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular items that happen to be used for the measurement); (b) non-crossing person response functions (i.e., a more able person must always have a better chance of success on an item than a less able person; (c) person-invariant calibration of test items (i.e., the calibration of the items must be independent of the particular persons used for calibration); (d) non-crossing item response functions (i.e., any person must have a better chance of success on an easy item than on a more difficult item); and (e) variable map (i.e., items and person must be simultaneously located on a single underlying latent variable). Based upon the difference in item and person locations on the variable map, items can be evaluated for their usefulness in providing information about persons' varying achievement levels. The benefit of Rasch approaches to measurement and construct modeling is the strong requirement that a set of items being used can measure a single construct (i.e., latent trait), the local independence of items, and sample-independent estimations of person and item parameters (i.e., invariant measurement).

On the occasion that raters facilitate the assessment process, the Many Facet Rasch (MFR) model can be utilized to simultaneously define person ability, item difficulty, and rater severity (Linacre, 1989/1994). The MFR model stems from the family of Rasch measurement models and can be used for both dichotomous or polytomous items (Wright & Mok, 2004). For polytomous items, the rating scale (RS) formulation of the model specifies consistent distances between rating scale categories across items (Wright & Masters, 1982). As pointed out by Engelhard (2013), the five requirements for invariant measurement can be extended to the context in which assessments are mediated by raters: (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) non-crossing 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); (d) non-crossing 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). When the data fit the requirements of the Rasch model, then it becomes possible to support invariant measurement that also implies rater-invariant measurement of performances (Engelhard, 2013).

Method

Scale, Items, and Judging Plan

Scale items (N = 22) were gleaned from the Jazz Big Band Performance Rating Scale (Wesolowski, 2015a). Each item was paired with a four-point Likert scale. Responses included "strongly agree," "agree," "disagree," and "strongly disagree." Expert judges (N = 23) were solicited based upon performance and teaching experience within the jazz idiom. Each judge directs a collegiate-level big band, holds at minimum a Master's degree in music, and is a certified adjudicator by his/her respective state music educator association. Judges were supplied with four anonymous recordings (middle school, n = 1; high school, n = 1, collegiate, n = 1; professional, n = 1). Middle school and high school recordings were gathered from district and state music performance assessments in the state of Florida. Collegiate and professional recordings were gathered from live performances in the states of Florida and Texas. The author and two outside evaluators carefully screened all performances utilized in this study for audible clarity. The recordings consisted of full performances in a medium swing style. Each judge was instructed to evaluate four performances to the best of their ability using the provided rating scale.

The judging plan was based upon the recommendations of Wright and Stone (1979) and Linacre and Wright (2004). In particular, the observations were organized in a manner whereby every parameter was linked to every other parameter both directly and indirectly. Specifically, each judge evaluated four recordings with two evaluations linked to another judge (e.g., Judge 1 evaluated recordings 1, 2, 3, 4; Judge 2 evaluated recordings 3, 4, 5, 6, etc.). Additionally, one common rater evaluated all 23 recordings. This plan allows all performance measures and item calibrations to be placed on one common scale by being compared directly and unambiguously thereby providing precise and accurate measures of all parameters (Linacre & Wright, 2004).

Data Analysis Procedures

In this study, the MFR model was used to explore the quality of ratings assigned to musical performances. Table 1 presents a set of statistics and displays based on the MFR model that can be used to examine the psychometric quality of musical performance assessments. The descriptions provided in Table 1 are adapted from Engelhard (2013), who illustrated the use of the MFR model to examine rating quality within the context of writing assessment. Although it is possible to calculate other indicators of rating quality based on the MFR model (e.g., interaction effects, see Engelhard, 2013), this study focuses on indices within three major categories: (a) logit-scale locations, (b) separation, and (c) model-data fit. In this section, a formulation of the MFR model is presented that can be used to examine the quality of music performance assessments. Then, the indices and displays from Table 1 are described. Results from

Table 1. Rating Quality Indices and Displays Based on the MFR Model for Musical Performance Assessments

Category	Indicators and Displays based on the MFR Model	Substantive Interpretation (Question)			
		Performance Facet	**Rater Facet**	**Item Facet**	**School Level Facet**
A. Logit-Scale Locations	1. Variable map	Where are the performances located on the construct being measured (musical achievement)?	Where are the raters located on the construct being measured (musical achievement)?	Where are the items located on the construct being measured (musical achievement)?	Where are the school levels located on the construct being measured (musical achievement)?
	2. Location of elements within the facet	What is the location of each performance (achievement)?	What is the location of each rater (severity/leniency)?	What is the location of each item (difficulty)?	What is the location of each school level (overall achievement)?
	3. Standard error	How precisely has the location of each performance been estimated?	How precisely has the location of each rater been estimated?	How precisely has the location of each item been estimated?	How precisely has the location of each school level been estimated?
B. Separation	4. Reliability of separation statistic	How spread out are the judged performance locations on the logit scale?	How spread out are the rater locations on the logit scale?	How spread out are the item locations on the logit scale?	How spread out are the school level locations on the logit scale?
	5. Chi-square statistic	Are the overall differences between performance locations significant?	Are the overall differences between rater locations significant?	Are the overall differences between item locations significant?	Are the overall differences between school level locations significant?
C. Model-data fit	6. Mean Square Error (*MSE*) and standardized fit statistics	How consistently has each performance been interpreted by the raters?	How consistently has each rater interpreted the items and rating scale categories across performances?	How consistently has each item been interpreted by the raters?	How consistently has each school level been interpreted by the raters?

Note. Adapted from Engelhard (2013).

the music performance assessment data are presented and interpreted in the subsequent section.

Many-Facet Rasch Model for Musical Performances

The Rasch-based statistics and displays for evaluating the quality of music performance assessments explored in this study were calculated using the Facets computer program (Linacre, 2014). Specifically, a Rating Scale (RS) formulation of the MFR model was used to explore the psychometric quality of the ratings examined in this study (Wright & Masters, 1982; Linacre, 1989/1994). The RS formulation of the MFR model used in this study included facets for ensemble performances, raters, items, and school levels. The model was specified as follows:

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \gamma_m - \tau_k \ , \qquad (1)$$

where

ln [$P_{nijmk}/P_{nijmk\text{-}1}$] = 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 = the logit-scale location (e.g., achievement) of performance *n*;
λ_i = the logit-scale location (e.g., severity) of rater *i*;
δ_j = the logit-scale location (e.g., difficulty) of item *j*;
γ_m = the logit-scale location (e.g., achievement) of school level *m*;
τ_k = the location on the logit scale where rating scale categories *k* and *k*-1 are equally probable.

The first four terms on the right side of the equation represent different facets of the rater-mediated assessment: Performances (θ), Raters, (λ), Items (δ), and School Levels (γ). The final term (τ) represents the difficulty associated with moving between categories *k* and *k*-1 on the logit scale (i.e., the rating scale category threshold). In this study, results from the MFR model shown in Equation 1 were used to explore the psychometric quality of a rater-mediated music performance assessment using Rasch-based statistics and displays.

Category A: Logit-scale locations. The first category of indices and displays based on the MFR model is *logit scale locations*. In the context of a rater-mediated music performance assessment, these indices provide a method for summarizing ensemble achievement, rater severity, and item difficulty on a single linear scale that represents the latent construct. As shown in Table 1, there are three indices within this category: (a) variable map; (b) calibration and location of elements within a facet; and (c) standard error.

Variable map. Once the MFR model has been applied to rating data, logit-scale locations are calculated for each facet. Because the Rasch model is unidimensional, it is possible to display the location estimates for each facet on a single linear scale. The variable map is a useful method for visually displaying descriptions of students, items, and other facets in terms of a latent variable. The usefulness of the variable map is a major factor in the adoption of Rasch modeling by many national and international assessments, including the National Assessment of Educational Progress (NAEP, 2009), the Program for International Student Assessment (OECD, 2009), and in music, the Model Cornerstone Assessments (NAfME, 2015). In the context of this study, the variable map provides a graphical representation of the ensemble performance, rater, item, and school level facets on a common "ruler" that represents musical achievement the logit scale. When adequate model-data fit is observed (described below), the variable map can be used as an operational definition of the latent construct.

Location of elements within a facet. It is informative to examine the estimates of logit-scale locations for individual elements within facets. In the context of this study, elements within facets include individual performances, raters, items, and school levels. These logit-scale locations correspond to the values that are plotted on the variable map. Higher values on the logit scale reflect higher musical achievement, more-severe rater judgments, and more-difficult items, while lower values on the logit scale reflect lower musical achievement, more-lenient rater judgments, and less-difficult items.

Standard error. In contrast to CTT, where an overall estimate of measurement error for an assessment is provided, the Rasch model provides standard error (*SE*) estimates for each element within a facet (e.g., each performance, rater, item, and school level). These estimates describe the range within which the element would be expected to fall if there were no measurement error. Smaller values of *SE* indicate more precise estimates, such that the logit-scale locations would be expected to remain stable across repeated administrations of an assessment.

Category B: Separation. The second category of indices and displays based on the MFR model is *separation.* After the logit-scale locations are estimated for each facet, it is possible to explore the degree to which the elements within a facet can be reliably differentiated from one another. As shown in Table 2, two statistics based on the Rasch model are used in this study to explore separation: (a) the reliability of separation statistic; and (b) a chi-square statistic.

Reliability of separation. The reliability of separation statistic (*Rel*) based on the Rasch model provides a method for examining the overall spread on the logit scale of elements within a facet. The interpretation of *Rel* varies depending on the facet to which it is applied. For the object of measurement (in the case of this study, ensemble performances), reliability of separation is interpreted in a comparable fashion to Cronbach's alpha coefficient, as it reflects a ratio of true-score to observed-score

variance. For other facets, the reliability of separation statistic describes the spread of differences in rater severity, item difficulty, and achievement within school levels.

Table 2. Summary Statistics from MFR Model

	Facets			
	Performance (θ)	**Rater** (λ)	**School Level** (γ)	**Item** (δ)
Logit-Scale Location				
M	0.38	0.00	0.00	0.00
SD	0.31	0.37	1.63	0.48
N	23	24	4	22
Infit MSE				
M	1.00	1.00	1.00	1.00
SD	0.16	0.18	0.07	0.20
Std. Infit				
M	-0.10	0.00	-0.10	-0.10
SD	1.20	1.30	1.40	1.50
Outfit MSE				
M	1.00	1.01	1.01	1.00
SD	0.17	0.19	0.08	0.20
Std. Outfit				
M	0.00	0.00	0.10	0.00
SD	1.30	1.40	1.40	1.60
Separation Statistics				
Reliability of Separation	0.75	0.79	>0.99	0.90
Chi-Square	87.2*	108.4*	2153.0*	210.7*
Degrees of Freedom	22	23	3	21

* $p < 0.05$

Chi-square statistic. A chi-square statistic (χ^2) can be calculated to determine whether the differences among logit-scale locations for elements of each facet are statistically significant.

Category C: Model-data fit. The third category of measurement quality indicators based on the MFR model is model-data fit. Model-data fit indices describe how closely empirical observations approximate the useful properties of the Rasch model. Evidence for adequate model-data fit supports the hypothesis of invariant

measurement. The approach to model-data fit analysis within Rasch measurement theory typically focuses on fit statistics that summarize residuals, or differences between model expectations and empirical observations. As shown in Tables 3-6, model-data fit is explored for each facet using statistical summaries of residuals: Infit and Outfit Mean Square Error (MSE) statistics.

Infit and outfit statistics. Infit and Outfit statistics are routinely used in Rasch analyses to explore model-data fit for facets related to persons and items. In the context of a rater-mediated musical performance assessment, these statistics can be calculated for facets related to performances, raters, items, and school levels.

The first step in calculating model-data fit statistics is to compare observed responses to the expected responses based on the Rasch model:

$$Y_{nij} = X_{nij} - P_{nij} \tag{2}$$

where

X_{nij} = observed response for Performance n scored by rater i on item j, and
P_{nij} = expected response probability for Performance n scored by rater i on item j, based on the Rasch model (Equation 1).

Next, the residuals (Y_{nij}) are standardized, and summarized as fit statistics that describe the degree to which adherence to the requirements for invariant measurement is observed in a set of data.

The Outfit *MSE* statistic is calculated by summing standardized residual variance across facets. Because it is unweighted, the Outfit *MSE* statistic is sensitive to "outliers," or extreme unexpected observations. Infit *MSE* statistics are also useful for evaluating model-data fit. However, they are less sensitive to outlying data because the residuals are weighted by the variance of an individual facet, which reduces the impact of unexpected observations. Similar to Outfit *MSE*, Infit *MSE* can be calculated for person- and rater-related facets. Because the exact sampling distribution for these fit statistics is not known (Wright and Masters, 1982; Engelhard, 2013), rules-of-thumb have been proposed for interpreting their values as they apply to specific types of facets, such as raters and students. Engelhard (2009) describes an acceptable range of Infit and Outfit *MSE* statistics of about 0.80 to 1.20. Values that are lower than about 0.80 suggest less variation in responses than expected, and values that are higher than about 1.20 suggest more variation in responses than expected; extreme values in both directions warrant further investigation. It is also possible to calculate standardized versions of Infit and Outfit using a cube root transformation of the MSE statistics, such that the values of standardized Infit and Outfit range from positive to negative infinity and follow a normal distribution. When data fit the model, the expected value for standardized Outfit statistics is 0.00, with a standard deviation of 1.00. Because the standardized fit statistics are approximate *t*-statistics, the critical values of +/- 2.00 are often used as critical values to indicate model-data fit (deAyala, 2008).

Interpreting fit statistics for raters. Engelhard (1994) described the application of Infit and Outfit statistics to rater-mediated writing assessment as a method for identifying idiosyncratic use of a rating scale. Specifically, low values of fit statistics may suggest that raters are not making full use of a rating scale (i.e., response sets, central tendency), or that there are dependencies across the ratings assigned to a group of students (i.e., halo error, score range restriction). On the other hand, high values of fit statistics may suggest haphazard ratings, or erratic use of the rating scale.

1. Good overall balance between winds and rhythm section	SD D A SA
2. Ensemble plays with a balanced sound in full passages	SD D A SA
3. Ensemble is balanced to the lead trumpet player during ensemble passages	SD D A SA
4. Ensemble plays with a large, full sound	SD D A SA
5. Ensemble accents figures in an appropriate manner	SD D A SA
6. Eighth note values are given appropriate duration	SD D A SA
7. Dynamic extremes are controlled	SD D A SA
8. Articulations are consistent with a good concept of jazz phrasing	SD D A SA
9. Ensemble maintains a steady time feel	SD D A SA
10. The rhythm section and winds share a common feel for the pulse	SD D A SA
11. Ensemble demonstrates a uniform feeling of pulse	SD D A SA
12. A steady tempo was kept throughout the performance	SD D A SA
13. Good overall blend between brass and saxophones	SD D A SA
14. Background figures are well-balanced to the soloist during solo sections	SD D A SA
15. Rhythm section makes appropriate balance adjustments between ensemble and solo sections	SD D A SA
16. Lead players perform with appropriate and idiomatic nuances	SD D A SA
17. Ensemble performs composition at an appropriate, idiomatic tempo	SD D A SA
18. Ensemble performs with a time feel appropriate to the composition	SD D A SA
19. Ensemble demonstrates a good concept of jazz phrasing	SD D A SA
20. Phrasing of eighth note lines is executed smoothly	SD D A SA
21. Ensemble performs with understanding of the swing eighth note concept	SD D A SA
22. Melodic lines end with an appropriate amount of emphasis	SD D A SA

Figure 1. 22-Item Jazz Big Band Performance Rating Scale (Wesolowski, 2015a

Results

Table 2 presents summary statistics from the MFR model facet analyses of music performance (n = 23), rater (n = 24), school level (n = 4), and scale items (n = 22). As shown in the table, the analysis indicated overall significant differences between performances ($\chi 2_{(22)} = 87.2,\ p < .05$), raters ($\chi 2_{(23)} = 108.4,\ p < .05$), school level ($\chi 2_{(3)} = 2153.0,\ p < .05$), and items ($\chi 2_{(21)} = 210.7,\ p < .05$). Acceptable reliability of separation for performance (Rel = .75) is comparable to Cronbach's coefficient alpha.

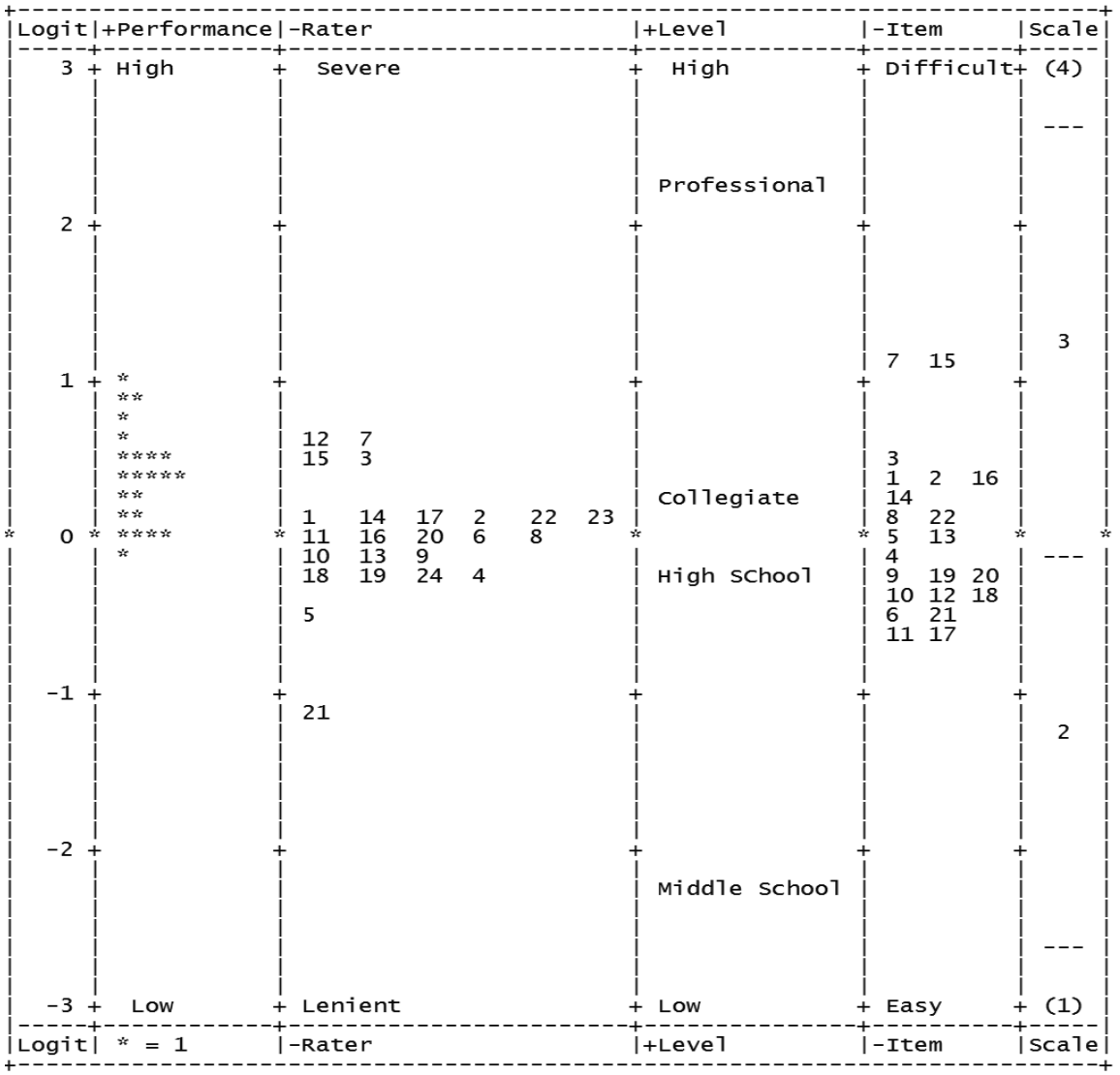

Figure 2. Variable Map

Acceptable reliability of separation for raters (*Rel* = .79) and high reliabilities of separation for school (*Rel* > .99) and item (*Rel* = .90) indicate a spread of the elements within each of the facets along the latent variable (e.g., ensemble performance achievement). Mean Infit and Outfit MSE values close to the expected value of 1.00 indicates good model fit and therefore provides validity evidence that the variable map (see Figure 1) can be utilized as a valid graphical representation of the model.

Figure 2 is a variable map that is a graphical display of the latent variable being investigated in this study. Specifically, the map contains the calibrations of facets included in the model on the same linear scale. In this study, the facets included ensemble performance, raters, school level, scale items, and rating scale categories. The labels at the top of the map indicate the facet and directionality of the measure. Column 1 includes the units of the logit scale whereby all facets can be calibrated and compared.

Column 2 includes the spread of ensemble measures (e.g., performance achievement), where each asterisk represents one ensemble performance (See Table 3). Performance achievement ranged from 1.05 to -0.08 logits (M = 0.38, SD = 0.31, N= 23). Higher measures indicate higher performance achievement. Evidence of misfit items is based upon Infit and Outfit MSE statistics outside of the rule-of-thumb ranges of 0.80 and 1.2 logits as indicated by Engelhard (2009) and standardized ranges of +/- 2.00 logits as indicated by deAyala (2008). Misfit performances included performances 4, 7, and 15. The results provide a fair and impartial rank ordering of ensembles based upon probabilistic distributions of responses as a logistic function of person and item parameters.

Column 3 represents the calibration of the rater measures (e.g., rater severity) (See Table 4). In order to define a frame of reference for interpreting the location of performance locations, all of the facets except the object of measurement are centered on the logit scale (mean set to zero). As a result, column three indicates that the average rater measures were anchored at 0.00 logits. Rater severity ranged from 0.67 logits (Rater 7 most lenient) to -1.08 logits (Rater 21 most severe) (SD = 0.38, N = 23). The directionality for the rater facet is negative indicating that a higher score yields a more lenient rater. Five raters demonstrated misfit rater patterns. Raters 11, 18, and 19 demonstrated muted rating patterns and raters 10 and 21 demonstrated haphazard rating patterns. As an example, an evaluation of rater 18's observed scores for performance 49 indicated the following: "strongly agree" (n = 17), "agree" (n = 5), "disagree" (n = 0), "strongly disagree" (n = 0). The muted rating pattern can be evidenced by an overwhelming use of the "strongly agree" category and no use of "disagree" and "strongly disagree." This may be interpreted as a possible halo error, where the rater fails to distinguish between conceptually distinct and independent aspects of ensemble performances (Engelhard, 2013). Rater 18's responses of the sample ensembles were similar for performances 9, 25, and 23. An evaluation of rater 11's observed scores for performance 4 indicated the following: "strongly agree" (n = 0), "agree" (n = 13), "disagree" (n = 9), "strongly disagree" (n = 1). Because rater 11's observed scores demonstrated similar tendencies for performances 6, 25, and 45, this may be interpreted as a response set error, where the rater interpreted and used the

rating scale categories in an idiosyncratic fashion. More specifically, rater 11's overuse of the middle categories (central tendency rater error) for performance 4 may not have been warranted by the ensemble's performance. Similar interpretations can be seen with rater 19.

Table 3. Calibration of the Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
4	2.67	1.05	0.16	1.42	2.84	1.50	3.27
17	2.54	0.93	0.16	1.07	0.61	1.07	0.56
12	3.16	0.88	0.17	1.05	0.44	1.18	1.32
9	2.05	0.78	0.15	0.85	-1.21	0.86	-1.16
22	2.85	0.59	0.16	1.24	1.74	1.24	1.74
5	3.05	0.51	0.16	0.96	-0.28	0.99	-0.06
19	2.47	0.48	0.16	0.93	-0.51	0.91	-0.66
2	2.31	0.47	0.15	0.98	-0.11	0.99	-0.04
16	2.53	0.45	0.15	0.84	-1.29	0.84	-1.26
10	2.34	0.44	0.15	1.02	0.18	1.03	0.24
1	2.55	0.40	0.16	0.99	-0.03	1.00	0.02
11	2.85	0.40	0.16	0.87	-1.05	0.82	-1.46
6	2.63	0.34	0.15	0.83	-1.29	0.83	-1.32
18	2.82	0.33	0.16	0.94	-0.44	0.93	-0.57
8	2.48	0.25	0.15	1.05	0.40	1.05	0.40
13	2.39	0.21	0.16	1.00	0.00	1.06	0.51
20	2.89	0.13	0.16	1.09	0.75	1.05	0.45
23	2.48	0.09	0.15	1.01	0.10	1.01	0.09
15	2.35	0.05	0.16	1.35	2.57	1.30	2.29
21	2.70	0.04	0.16	0.99	-0.02	0.98	-0.16
7	2.66	0.03	0.16	0.77	-1.99	0.77	-1.97
14	2.58	-0.04	0.16	0.89	-0.86	0.88	-0.97
3	2.58	-0.08	0.15	0.79	-1.70	0.80	-1.63
Mean	2.61	0.38	0.16	1.00	-0.05	1.00	-0.02
SD	0.25	0.31	0.01	0.16	1.22	0.17	1.29

Note. The performances are presented in Measure order, from high to low.

The fourth column includes the calibrations used to represent the performance achievement level of the ensembles (See Table 5). The measures for the school level facet were anchored at 0.00 logits. Middle school ensembles scored the lowest (M =-2.28, SE = 0.06), followed by high school (M = -0.24, SE = 0.06), collegiate (M = 0.22, SE = 0.06), and professional (M = 2.30, SE = 0.08). Because the level facet is related to performance ordering, the directionality is positive, such that higher measures indicate higher levels

of achievement. The ordering of the ensembles from low to high as logically expected based upon their categorization further validates the measure.

Table 4. Calibration of the Rater Facet

Rater ID	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
7	2.42	0.67	0.17	0.94	-0.37	0.95	-0.34
12	2.44	0.64	0.17	0.92	-0.57	0.92	-0.55
3	2.27	0.54	0.17	1.09	0.68	1.09	0.63
15	2.41	0.50	0.18	1.07	0.50	1.05	0.39
23	2.61	0.18	0.18	0.89	-0.75	0.89	-0.75
14	2.56	0.16	0.18	1.13	0.91	1.18	1.28
1	2.60	0.11	0.07	0.96	-0.66	0.95	-0.83
2	2.56	0.11	0.18	1.14	1.00	1.15	1.07
17	2.56	0.09	0.17	0.84	-1.13	0.85	-1.10
22	2.67	0.08	0.17	1.01	0.15	1.02	0.17
16	2.58	0.04	0.18	0.74	-1.98	0.74	-1.98
20	2.45	0.01	0.17	0.96	-0.21	0.95	-0.29
8	2.66	-0.02	0.18	1.20	1.34	1.20	1.39
6	2.70	-0.03	0.18	1.06	0.43	1.05	0.40
11	2.70	-0.06	0.17	0.71	-2.18	0.71	-2.25
9	2.58	-0.10	0.17	0.97	-0.17	0.95	-0.35
13	2.68	-0.11	0.18	1.07	0.53	1.03	0.25
10	2.62	-0.12	0.17	1.40	2.58	1.36	2.36
18	2.65	-0.25	0.17	0.69	-2.39	0.70	-2.36
24	2.86	-0.26	0.18	1.10	0.72	1.27	1.68
19	2.51	-0.30	0.18	0.72	-2.19	0.71	-2.31
4	2.70	-0.31	0.18	1.09	0.65	1.27	1.72
5	2.86	-0.49	0.18	1.06	0.46	1.02	0.20
21	2.89	-1.08	0.18	1.31	2.06	1.31	2.06
Mean	2.61	0.00	0.17	1.00	-0.02	1.01	0.02
SD	0.15	0.38	0.02	0.18	1.30	0.19	1.38

Note. The raters are presented in Measure order, from high (severe) to low (lenient).

Column 5 represents the item calibrations (See Table 6). Item calibrations were anchored at 0.00 logits and the directionality is negative, associating higher scores with easier items. Calibrations ranged from 1.10 logits (Item 15 as easiest item) to -0.59 logits (Item 11 as most difficult item) (*SD* = 0.47). No misfit items were evidenced based on the Infit and Outfit MSE or standardized fit statistics.

The sixth column represents the rating scale structure. The response format included "strongly disagree," "disagree," "agree," and "strongly agree." The categories are represented in column 6 of the variable map. Inspection the rating scale structure

based upon the quantitative guidelines set forth by Linacre (1999, 2002) indicates cooperation of the rating scale categories to produce meaningful measures.

Table 5. Calibration of the School Level Facet

Level	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
Middle School	1.78	-2.28	0.06	0.98	-0.46	0.98	-0.47
High School	2.55	-0.24	0.06	1.09	1.72	1.09	1.70
Collegiate	2.77	0.22	0.06	0.89	-2.08	0.90	-1.96
Professional	3.47	2.30	0.08	1.03	0.55	1.06	1.06
Mean	2.64	0.00	0.07	1.00	-0.07	1.01	0.08
SD	0.60	1.63	0.01	0.07	1.39	0.07	1.42

Note. The levels are presented in Measure order, from high to low.

Discussion

This study focused on the description of a new method to examine rater quality in rater-mediated assessments of music ensemble performances based upon the Many Facet Rasch (MFR) model. Although large ensemble performance evaluations are not considered high stakes educational assessments, equity in the evaluation process is important as resulting scores often inform perceptions of program quality and instructor effectiveness (Boyle, 1992). Furthermore, suggestions have been offered to consider large group performance assessments as a means to measure teacher effectiveness and student growth in music performance because large group performance assessments, "like standardized tests- provide a third-party evaluation consisting of numerical scores that can be used to compare the achievement of one group or director to that of another." (Hash, 2013, p. 163). Federal priorities, however, specify that suitable measures for deriving student growth data should be rigorous and be comparable across classrooms (Wesolowski, 2014; Wesolowski, 2015b, NCCTQ, 2011). Research related to adjudicator reliability, validity, fairness, and scales developed utilizing CTT measurement models are inadequate for these purposes due to sample and test dependency of estimated person parameters and item parameters. CTT models limit the ability to develop valid and reliable measures and to make informed inferences related to examinee ability and item difficulty that extend beyond the context of a single assessment situation. Additionally, in instances where raw score data are utilized, linear magnitudes are often assumed. Raw scores, however, are in fact ordinal counts of observations that indicate "more" or "less" (Wright & Masters, 1982; Saltzberger, 2010). As Bond and Fox (2007) state:

> In terms of Stevens's (1946) levels… nominal and ordinal levels are NOT any form of measurement in and of themselves. Admittedly, we concur that his

> interval and ratio levels actually would constitute genuine measurement, but the scales to which we routinely ascribe that measurement status in the human sciences are merely *presumed* to have measurement properties; those measurement properties are almost never tested empirically. It is not good enough to allocate numbers to human behaviours and then, merely to *assert* that this is measurement in the social sciences. (p. 4)

Table 6. Calibration of the Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
15	2.19	1.10	0.15	1.09	0.73	1.12	0.97
7	2.20	1.08	0.15	1.12	0.95	1.09	0.76
3	2.43	0.46	0.15	1.39	2.91	1.37	2.76
16	2.45	0.41	0.15	0.79	-1.83	0.81	-1.65
1	2.47	0.36	0.15	1.08	0.70	1.07	0.58
2	2.48	0.34	0.15	0.87	-1.09	0.86	-1.17
14	2.50	0.27	0.15	1.08	0.66	1.07	0.58
8	2.56	0.14	0.15	0.67	-3.01	0.69	-2.79
22	2.58	0.07	0.15	0.81	-1.62	0.82	-1.51
5	2.60	0.02	0.15	0.90	-0.83	0.91	-0.76
13	2.63	-0.05	0.15	1.18	1.41	1.28	2.13
4	2.67	-0.17	0.15	0.96	-0.28	0.97	-0.20
9	2.68	-0.19	0.15	1.21	1.60	1.21	1.63
20	2.69	-0.21	0.15	0.99	-0.06	0.99	0.00
19	2.70	-0.24	0.15	0.75	-2.12	0.77	-2.03
18	2.75	-0.38	0.15	0.93	-0.52	0.93	-0.56
10	2.76	-0.40	0.15	0.84	-1.33	0.85	-1.22
12	2.77	-0.42	0.15	1.10	0.83	1.09	0.77
6	2.79	-0.50	0.15	0.97	-0.22	0.96	-0.25
21	2.81	-0.54	0.16	0.76	-1.96	0.76	-2.05
17	2.82	-0.57	0.16	1.44	3.07	1.51	3.51
11	2.83	-0.59	0.16	0.97	-0.15	0.94	-0.45
Mean	2.61	0.00	0.15	1.00	-0.10	1.00	-0.04
SD	0.18	0.47	0.00	0.20	1.55	0.20	1.58

Note. The items are presented in Measure order, from high (difficult) to low (easy).

Wright and Masters (1982) further indicate:

> For observations to be combined into measures they must be brought together and connected to the idea of measurement which they are intended to imply. The recipe for bringing them together is a mathematical formulation or measurement

> model in which observations and our ideas about the relative strengths of persons and items are connected to one another. (p. 4)

An important advantage of the Rasch measurement model over other measurement models is the linear logistic transformation of ordinal-level raw scores into interval-scaled measures with strict requirements for invariant measurement (Saltzberg, 2010).

As brought to light in this study, measurement of music ensemble performances includes many factors that can contribute to the variability of observed scores. These factors include ensemble ability level, the difficulty of the task, the severity of the raters, and the structure of the rating scale. The utilization of Rasch measurement theory, and more specifically, the MFR model, provides a framework for obtaining objective linear measurements of ensemble performance achievement that are invariant over these factors. The scaled scores take into account each of the factors in order for all ensembles to be directly compared. Adjustments to these factors can improve objectivity and fairness in the evaluation process by minimizing measurement error.

In the case of rater-mediated assessments, the evaluation of fit indices and observed scores can detect a wide range of rater effects in addition to rater severity. These can include rater accuracy, halo effects, central tendencies, and restrictions of range (Engelhard, 1994; Saal, Downey, & Lahey 1980). An area for future investigation is a thorough analysis of rater error and sampling distributions of observed scores in music performance assessments. A better understanding of specific rater effects may improve the accuracy of future evaluations and rater-training processes. More specifically, indices of rater training can be used to monitor the quality of raters over time, provide feedback to raters, screen out inaccurate raters, and evaluate the effects of rater training programs. The application of the methods demonstrated in this paper can provide quality control to inform training processes, minimize potential raters errors, and contribute to the overall improvement of rating scale development.

In summary, the application of the MFR model to ensemble performances proves to be a fruitful method for investigating rater behavior in the evaluation of musical performances. The implementation of rigorous measurement models such as the MFR-RS model to music assessment research provides an approach for developing linear, unidimensional measures from ordinal observations thereby establishing equity and fairness to the music performance assessment process.

References

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education, 21*, 246-255.

Brennan, R. L. (2001). Generalizability theory: Statistics for social science and public policy. New York: Springer-Verlag.de Ayala, R. J. (2009) *The theory and practice of item response theory.* New York: Guilford Press.

Asmus, E. P. (1985). The development of a multidimensional instrument for the measurement of affective responses to music. *Psychology of Music, 13*, 19-30.
Asmus, E. P. (1989). Factor analysis: A look at the technique through the data of rainbow. *Bulletin of the Council for Research in Music Education, 101*, 1-29.
Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking, *Language Testing, 12,* 238-257.
Barnes, G. V., & McCashlin, R. (2005). Practices and procedures in state adjudicated orchestra festivals. *Update Applications of Research in Music Education, 23*(2), 34-41.
Bergee, M. J. (1992). A scale assessing music student teachers' rehearsal effectiveness. *Journal of Research in Music Education, 40,* 5-13.
Bergee, M. J. (1995). Primary and higher-order factors in a scale assessing concert band performance. *Bulletin of the Council for Research in Music Education, 126*, 1-14.
Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education, 51,* 137-150.
Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55*, 344-358.
Bergee, M. J., & McWhirter, J. L. (2005). Selected influences on solo and small-ensemble festival ratings: Replication and extension. *Journal of Research in Music Education, 53*, 177-190.
Bergee, M. J., & Platt, M. C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 51*, 342-353.
Bergee, M. J., & Westfall, C. R. (2005). Stability of a model explaining selected extramusical influences on solo and small ensemble festival ratings. *Journal of Research in Music Education, 53,* 358-374.
Boeckman, J. (2002). Grade inflation in band contest ratings: A trend study. *Journal of Band Research, 38*(1), 25-36.
Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences, second edition.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.
Boyle, D. J. (1992). Program evaluation for secondary school music programs. *NASSAP Bulletin, 76*, 63-68.
Brakel, T. D. (2006). Inter-judge reliability of the Indiana State School Music Association high school instrumental festival. *Journal of Band Research, 42*(1), 59-69.
Brand, M. (1985). Development and validation of the home musical environment scale for use at the early elementary level. *Psychology of Music, 13*, 40-48.
Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.
Burnsed, V. Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research, 21*(1), 22-29.
Ciorba, C. R., & Smith, N. Y. (2009). Juries using a multidimensional assessment rubric measurement of instrumental and vocal undergraduate performance juries. *Journal of Research in Music Education, 57*, 5-15.

Colwell, R. (1970). The development of the music achievement test series. *Bulletin of the Council for Research in Music Education, 22,* 57-73.
Conrad, D. (2003). Judging the judges: Improving rater reliability at music contests. *NFHS Music Association Journal, 20*(2), 27-31.
DeAyala, R. J. (2008). *The theory and practice of item response theory: Methodology in the social sciences*. New York: Guliford Press.
Edwards, J. S., & Edwards, M. C. (1971). A scale to measure attitudes toward music. *Journal of Research in Music Education, 19,* 228-233.
Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.
Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (261-287). Mahwah, NJ: Erlbaum.
Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602.
Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
Engelhard, G., Jr., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research & Perspective, 9,* 40-45.
Fautley, M. (2010). *Assessment in music education*. New York: Oxford University Press.
Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23*, 186-189.
Fiske, H. E. (1983). *The effect of a training procedure in music performance evaluation on judge reliability.* Ontario Educational Research Council Report, Canada.
Fox, G. C. (1990). Making music festivals work. *Music Educators Journal, 76*(7), 59.
Garman, B. R., Boyle, J. D., & DeCarbo, N. J. (1991). Orchestra festival evaluations: Interjudge agreement and relationships between performance categories and final ratings. *Research Perspectives in Music Education, 2*, 19-24.
Harman, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press.
Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education, 60*(1), 81-100.
Hash, P. M. (2013). Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review, 114*(4), 163-169.
Jellison, J. (1985). An investigation of the factor structure of a scale for the measurement of children's attitudes toward handicapped peers within regular music environments. *Journal of Research in Music Education, 33,* 167-177.
Jorsekog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (47-77). Mahwah, NJ: Erlbaum.

King, S. E., & Burnsed, V. (2009). A study of the reliability of adjudicator ratings at the 2005 Virginia band and orchestra directors association state marching band festivals. *Journal of Band Research, 45*(1), 27-32.
Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education, 56*, 322-337.
Kruth, E. (1970). Adjudication and the music festival. *The Instrumentalist, 24*(6), 48.
Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement, fourth edition* (387-431). Westport, CT: American Council on Education and Praeger.
Latimer, M. E., Bergee, M. J., & Cohen, M. L. (2010). Performance assessment rubric reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education, 58*, 168-183.
Linacre, J. M. (1989/1994). *Many-facet Rasch measurement.* Chicago: MESA Press.
Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.
Linaacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.
Linacre, J. M. (2014). *Facets (Version 3.71.4)* [Computer software]. Chicago, IL: MESA Press.
Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many facet data. In E. V. Smith, Jr., & R. M. Smith (Eds*). Introduction to Rasch measurement: Theories, models, and applications* (296-321). Maple Grove, MN: JAM Press.
Morrison, S. J., Price, H. E., Geiger, C. G., & Cornacchio, R. A. (2009). The effect of conductor expressivity on ensemble performance evaluation. *Journal of Research in Music Education, 57*, 37-49.
NAEP: National Assessment for Educational Progress. (2009). *NAEP technical documentation*. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/
NAfME: The National Association for Music Education. (2015). *Music Model Cornerstone Assessment*.
NAfME: The National Association for Music Education. (n.d.). *National Music Adjudication Coalition Instrumental Jazz Ensemble Music Assessment Form.* Retrieved from http://musiced.nafme.org/files/2013/02/NMACinstjazzform.pdf
NCCTQ: The National Comprehensive Center for Teacher Quality. (2011). *Measuring teachers' contributions to student learning growth for non-tested grades and subjects.* Washington DC: Educational Testing Service.
Nichols, J. P. (1991). A factor-analysis approach to the development of a rating scale for snare drum performance. *Dialogue in Instrumental Music Education, 15*, 11-31.
Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education, 55*, 237-251.
OECD: Organization for Economic Co-operation and Development. (2009). *PISA data analysis manual: SAS, second edition*. PISA, OECD Publishing.

Price, H. E., & Chang, E. C. (2005). Conductor and ensemble performance expressivity and state festival ratings. *Journal of Research in Music Education, 53*, 66-77.
Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education, 24,* 119-128.
Rasch. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980 Chicago, IL: University of Chicago Press).
Russell, B. E. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education, 184*, 21-34.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.
Saltzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions, 24*(2), 1273-1275.
Sandstrom, G. M., & Russo, F. A. (2013). Absorption in music: Development of a scale to identify individuals with strong emotional responses to music. *Psychology of Music, 41,* 216-228.
Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education, 45*, 259-272.
Shaw, C. N., & Tomcala, M. (1976). Music attitude scale for use with upper elementary school children. *Journal of Research in Music Education, 24,* 73-80.
Silvey, B. A. (2009). The effects of band labels on evaluators' judgments of musical performance. *Update: Applications of Research in Music Education, 28*(1), 47-52.
Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education, 57*, 217-235.
Stevens, S. S. (1946). On a theory of scales of measurement. *Science, 103*, 667-680.
Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal, 101*, 77-85.
Wesolowski, B. C. (2015a). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music.* doi:10.1177/0305735614567700
Wesolowski, B. C. (2015b). Tracking student achievement in music performance: Developing student learning objectives for growth model assessments. *Music Educators Journal,* 102, 39-47.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.
Wright, B. D., & Mok, M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith and R. M. Smith (Eds.). *Introduction to Rasch Measurement* (1-24). JAM Press: Maple Grove, MN.
Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education, 50*, 245-255.