Examining Rater Precision in Music Performance Assessment: An Analysis of Rating Scale Structure Using the Multifaceted Rasch Partial Credit Model

Brian C. Wesolowski
The University of Georgia

Stefanie A. Wind
The University of Alabama

George Engelhard, Jr.
The University of Georgia

The use of raters as a methodological tool to detect significant differences in performances and as a means to evaluate music performance achievement is a solidly defended practice in music psychology, education, and performance science research. However, psychometric concerns exist in raters' precision in the use of task-specific scoring criteria. A methodology for managing rater quality in rater-mediated assessment practices has not been systematically developed in the field of music. The purpose of this study was to examine rater precision through the analysis of rating scale category structure across a set of raters and items within the context of large-group music performance assessment using a Multifaceted Rasch Partial Credit (MFR-PC) Measurement Model. Allowing for separate parameterization estimation of the rating scale for each rater can more clearly detect variability in rater judgment and improve model-data fit, thereby enhancing objectivity, fairness, and precision of rating quality in the music assessment process. Expert judges ($N = 23$) rated a set of four recordings by middle school, high school, collegiate, and professional jazz big bands. A single common expert rater evaluated all 24 jazz ensemble performances. The data suggest that raters significantly vary in severity, items significantly vary in difficulty, and rating scale category structure significantly varies across raters. Implications for the improvement and management of rater quality in music performance assessment are provided.

Received: February 22, 2015, accepted September 3, 2015.

Key words: assessment, judgment, precision, Rasch, raters

Value judgments and decision-making permeate our world, from seemingly casual, aesthetic value judgments of art (Chatterjee, Widick, Sternschein, Smith, & Bromberger, 2010), wine (Jackson, 2009), and film (Ginsburgh & Weyers, 2007), to more robust and systematic judgments related to interventions of at-risk school children (DiStefano, Greer, Kamphaus, & Brown, 2014), medical treatments (Campbell, Kolobe, Osten, Lenke, & Girolami, 1995), and high-stakes athletic events (Looney, 1997). The role of human reasoning and decision making has long held the attention of scholars in the fields of philosophy, natural teleology, and aesthetics as evidenced from the writings of Descartes and Cottingham (1986) and Kant and Meredith (1986) to more modern, empirical investigations into the epistemology and psychology of human judgment (Bishop & Trout, 2005). In Freudian psychoanalytic theory of human personality and judgment, the pleasure principle indicates that psychical activity is based upon rapid and subjective decision making between pleasure and pain, good and bad (Freud, 1920). Initial human decisions are made so rapidly that cognitive reason and logic do not have time to influence the reaction, and are made before our minds have had the chance to perceive what is being judged (Freeman, Stolier, Ingbretsen, & Hehman, 2014; Todorov, Said, Engell, & Oosterhof, 2008). In the evaluation of musical performances, raters are subjected to rapid, real-time decision-making processes based upon immediate reactions to trait inferences (Thompson, Williamon, & Valentine, 2007). As a result, music evaluation processes are dominated primarily by rater intuition and holistic paradigms (Davidson & Coimbra, 2001; Forbes, 1994; Mills, 1991; Stanley, Brooker, & Gilbert, 2002). Music performances, however, often need to be evaluated using objective, data-driven processes for the purpose of music research or the documentation of student achievement. Under these circumstances, music performance assessments are designed with the intent to systematically and objectively evaluate both processes of performance and products of performance.

Music Perception, volume 33, issue 5, pp. 662–678, issn 0730-7829, electronic issn 1533-8312. © 2016 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, http://www.ucpress.edu/journals.php?p=reprints. DOI: 10.1525/mp.2016.33.5.662

Rater precision plays an important role in the facilitation of fair rater-mediated assessment, yet it is an area that has not been investigated in music research under the umbrella of item response theory. Numerous studies have examined issues related to rater-centered variability in music performance adjudication research, with a focus toward factors such as purpose of the assessment, musical knowledge, extent of training, and personality (see McPherson & Schubert, 2004; McPherson & Thompson, 1998, for full review). As highlighted by Wesolowski, Wind, and Engelhard (2015), such studies lack the clarification of differential rater functioning and bias and therefore should be read critically. Additionally in music research, the validation of measures predicated on rater behavior and/or focus of rater behavior in of itself is overwhelmingly dominated by the use of inter- and intra-rater reliability (Bergee, 2003; Brakel, 2006; Fiske, 1977, 1979; Hash, 2012; Latimer, Bergee, & Cohen, 2010, for example).

In such instances, the correlational approach to internal consistency can obscure systematic differences in rater severity and provide a misleading result in the rankings of performance achievement or raters severity. Additional concerns related to differences in rater severity, use of the rating scale, and standard errors of measurement also contribute to rater precision, and thus warrant examination.

The purpose of this study was to examine rater precision across a set of raters and items within the context of large-group music performance assessment using a Multifaceted Rasch Partial Credit (MFR-PC) Measurement Model. More specifically, this paper investigated rater precision through: (a) the analysis of variability in rater behavior (e.g., severity and leniency); (b) raters' use of rating scale structures (i.e., category response functions) across individual raters; and (c) standard errors of measurement. This study was guided by the following research questions:

1. How do raters vary in leniency and severity?
2. How do items vary in difficulty?
3. How does the structure of the rating scale vary across individual raters?

PRECISION AND VALIDITY IN HUMAN JUDGMENT

The rating of performances is a complex cognitive process embedded with elaborate task demands that can often lead to systematic variance in ratings (Hamp-Lyons & Henning, 1991; Myford & Wolfe, 2004). According to Guliford (1936), "Raters are human and they are therefore subject to all the errors to which humankind must plead guilty" (p. 272). Statistical interpretations of Brunswik's (1952) lens model have often been used to study varying levels of achievement in human judgment (Karelaia & Hogarth, 2008). The fundamental premise of the lens model outlines concerns of rater precision and ecological validity in human decision-making. The lens model was brought to light by Hammond (1996), who highlighted two primary predictors that affect judgmental precision in the context of applied psychology: (a) predictability of the criterion based upon cues (i.e., how well the criterion function in relation to the precision of raters' use of rating scale categories); and (b) how well these cues match rater behavior and the environment (i.e., ecological validity of the measurement apparatus).

Karelaia and Hogarth's (2008) meta-analysis of studies implementing lens model indices from the Brunswikian tradition provide an in-depth overview of important task demands and factors that affect the precision of human decision making processes. In particular, three salient factors gleaned from Karelaia and Hogarth's investigation are relevant to this study: (a) linear models describing functional relations between criterion and cues are overall better than nonlinear models for the precision and accuracy of judgmental decisions; (b) equally rated criterion and equally cued weighting schemes are preferable for rater precision due to an imbalance between individual expectations and environmental structure, guaranteeing that the rater considers each criterion and cue fully and equally; and (c) level of expertise is preferable for meaningfulness of content, but not for accuracy and precision.

As pointed out by Engelhard (2013), the Brunswik lens model can bring to light three important considerations when using raters as a source for deriving observable qualities in psychological and behavoral research. First, rating criteria and rating scale structure serve as cues for which raters base their judgment. Therefore, these cues must be carefully analyzed for functionality, rater use, and rater precision. As demonstrated in this paper, the benefit of rescaling ordinal-level, observable rater data to a log odds interval scale under the strict requirements of invariant measurement is a fruitful method for evaluating rater precision in this capacity. Second, rater-mediated assessments and related rater behavior must be contextualized within specific assessment systems, as the lens model is grounded in ecological psychology. The approach of this study is contextualized by the use of a linear, unidimensional latent construct (i.e., jazz big band performance achievement) to evaluate rater precision and item functionality. Third, in considering rater precision, the lens model illuminates the need for a methodology to systematically manage potentially sporadic (i.e., too unpredictable) and/or muted (i.e., too predictable) rating results. As Brunswik (1952) notes, " . . . imperfections of [judgmental] achievement may in part be ascribable to the lens itself, that is, to the organism [i.e., rater] as an imperfect machine" (p. 23). The major benefit of the Rasch model is that when adequate fit to the model is observed, invariant measurement is achieved. In the application presented in this study, invariant measurement implies that the calibration of musical performances, calibration of raters, and calibration of items are independent and able to be systematically evaluated using fit indices, each with individual measures of standard error.

STUDIES IN RATER BEHAVIOR

Two broad categories of studies in rater behavior exist related to performance assessments. First is a rater behavior-centered approach that investigates ecological content of human judgment. These studies stress both rater attributes as well as characteristics of the environment. Examples include how rater variability is affected by rater experiences and training procedures (Barrett, 2001), prior knowledge and effects of feedback (Elder, Knoch, Barkhuizen, & von Randow, 2005), content knowledge and cognition (Freedman & Calfee, 1983), and ecological context (Hogarth, 1987). In music performance assessment research, process models have addressed rater-behavior centered attributes by considering non-musical factors such as rater personality, experience and musical ability, training in adjudication, familiarity with the performer and/or repertoire (see McPherson & Schubert, 2004, and McPherson & Thompson, 1998, for a more thorough review). There exists a connection between task demands and rater-centered experiences as outlined in the dual processing theory of human cognition due to two concurrent yet separate active systems of operations: (a) intuitive thought processes, and (b) reflective thought processes (Sun, 1994). As a result, the complex relationship between personal experience, expertise, and task demands set upon raters can influence the judging process (Platz & Kopiez, 2012). However, correlation does not equal causality when linking variability in judgments to rater errors and "biases" (Wesolowski et al., 2015). As demonstrated by Wesolowski et al. (2015), empirical evidence of differential rater functioning (DRF) (i.e., raters' statistically significant differential leniency/severity among subgroups) can be tested for their effects; however, the explanation for the phenomenon is qualitative and can only be interpreted through expert evaluation of systematic patterns of misfit. DRF uses a different set of statistical subroutines that can provide an entirely new story of patterns in rater behavior. DRF analysis was found to flag raters as evidencing differential severity or leniency that could have gone unnoticed without conducting the post hoc DRF analyses. Raters that fit the model demonstrated DRF. Without this analysis, these patterns would not have been identified. Raters may be precise in their estimates of individual true performances (i.e., good model data fit); however, systematic patterns emerged in their scoring when comparing the subgroups of performances. Some rater functioning, therefore, was systematically altered according to subgroup. Arguably in the previously presented models, both concepts have been treated similarly. The blurring of these two concepts has provided a distorted view of rater functioning and patterning when evaluating musical performances.

The second category is an empirically driven approach that includes investigation into rater effects, statistical indices, and quality of assigned ratings. According to Eckes (2008), rater variability under these conditions can stem from: (a) the degree to which raters comply with the measurement tool; (b) the way raters interpret criteria in operational scoring sessions; (c) the degree of leniency and severity exhibited; (d) raters' understanding of the measurement tool's rating scale categories; and (e) the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. Such studies include analyses of raters' use of rating scale structure, patterns of centrality, accuracy, and differential dimensionality based upon particular measurement methodologies. Applications of the Rasch family of measurement models to music performance assessment situations are providing improved methods for fair, valid, and reliable measures of achievement using raters where such variability can be detected, separated, and controlled for.

In educational contexts such as university music system juries and other related performances, raters are used in order to provide summative assessments of musical performances under carefully developed scoring schemes and written sets of performance criteria. These local conditions often using hermeneutic scoring system approaches where minor differences in rater scoring are often welcomed, as they contribute rich and varied content for the improvement of student performance as well as provide an opportunity to make curricular decisions explicitly reflect appropriate standards and criteria (Haswell, 2001; O'Neill, 2002). However, in more formalized standards-based performance assessments systems such as the Associated Board of the Royal Schools of Music's (ABRSM) performance assessments in the United Kingdom, the National Association for Music Education's (NAfME) new National Core Arts Standards performance assessments in the United

States, or the Australian Music Examination Board's (AMEB) practical exams in Australia, differences in psychometric scoring systems due to rater effects can be a source of potential error that obscure the validity of developed measures and the reliability/accuracy of the rater. Under these assessment conditions, raters are used for the purpose of providing evaluative feedback, but more importantly, for providing accurate and reliable scoring based upon standardized musical achievement levels through the use of pre-established criteria that is either grade- or age-level dependent. Stakeholders, administrators, and policy makers are becoming increasingly aware of rater effects as a source of error under formalized performance assessment conditions in the fields of writing, science, and medicine as results are often used for high stakes decision making and have a large impact on community, national, and international perceptions of program quality (Boyle, 1992). In music research however, attention to the effects of rater decisions on psychometric scoring systems is in its infancy, with most pioneering studies on adjudication bias, rater consistency, and rater consensus setting a precedent for the field with a focus on indices of rater agreement or reliability (Duerksen, 1972; Fiske, 1978; Flores & Ginsburgh, 1996, among many others).

Rater-mediated performance assessment is equally as important as a methodological tool in the research areas of music psychology and music performance science. As Thompson and Williamon (2003) note, "... no published work has specifically addressed the issue of how existing performance assessment protocols might be developed into reliable tools for researchers." (p. 22). Under research-based conditions, a rater's reliable and accurate detection of differences in performance achievement is an objectively meaningful method for providing significant evidence of treatment effects, efficacy of various intervention types, or overall differences in performance parameters that affect the listening experiences and judgmental processes. However solidly defended, use of rater-mediated data in experimental research is not without empirical concerns of reliability and consistency (Thompson & Williamon, 2003).

A concern of Thompson and Williamon are the subjective use of the forced-choice nature of items use. Specifically:

> ... experienced musicians may develop a kind of internal segmented marking scheme, perhaps one that is specific to the piece being performed ... Clearly ... the potential existence of highly personal category systems is problematic for the researcher faced with the challenge of reliably quantifying performance difference of change (pp. 26-27).

In these instances, music research has been limited by traditional psychometric approaches to measurement (e.g., Classical test theory) through the application of consensus and consistency estimates to directly evaluate rater effectiveness. However, Bock and Jones (1968) note, "In a well-developed science, measurement can be made to yield invariant results over a variety of measurement models and over a range of experimental conditions for any one method" (p. 9).

CONCERNS WITH TRADITIONAL PSYCHOMETRIC APPROACHES TO RATER-MEDIATED MUSIC ASSESSMENT

Variability in inter- and intra-rater agreement at the item level is a product of the fields' acceptance for "what you see is what you get," where results of an assessment situation is based on "luck of the rater draw" (Engelhard, 2013). This dissenting perspective of fair evaluation practices using structured assessment schemes is also consequence and frustration of the limitations of Classical Test Theory (CTT) as a means for predicting outcomes of psychological testing. CTT is often the primary measurement model for evaluating consistency estimates and consensus estimates in music performance assessment (Bergee, 2003; Brakel, 2006; Burnsed, Hinkle, & King, 1985; Conrad, 2003; Fiske, 1983; Hash, 2012; King & Burnsed, 2007; Norris & Borst, 2007; Silvey, 2009, for example) and is an accepted methodology in music perception and human response for analyzing rater behavior under various research conditions (Aruffo, Goldstone, & Learn, 2014; Castro & Lima, 2014; Hutchins, Hutka, & Moreno, 2014; Labbe & Grandjean, 2014, for example). CTT methodology "models the statistical nature of the scores and focuses attention on the consistency of results from the [measurement] instrument (i.e., reliability)" (Wilson, 2005, p. 88). Consensus estimates of inter-rater reliability are based upon the assumption that absolute agreement between raters can be achieved on how they apply the various levels of scoring to the observed behaviors being rated. According to Stemler (2004), using consensus estimates as a method for monitoring rater quality can be imprecise, misleading, and disadvantageous for four reasons: (a) inter-rater reliability statistics must be computed separately for each pair of judges and each item; (b) cost and time efficiency in scale development and the training of judges to come to an exact agreement; (c) forcing judges into exact agreement can threaten construct validity by reducing statistical independence of the ratings (Linacre, 2002); and (d) estimates can be overly conservative. According to Henning (1997):

> . . . two raters may agree in their score assignments and both be wrong in their judgments simultaneously in the same direction, whether by overestimating or underestimating true ability. If this happens, then we have a situation in which raters agree, but assessment is not accurate or reliable because the ratings fail to provide an accurate approximate approximation of the true ability score. Similarly, it is possible that two raters may disagree by committing counterbalancing errors in opposite directions; that is where one rater overestimates true ability, and the other rater underestimates true ability. In this latter situation, it happens that the average of the two raters' scores may be an accurate and reliable reflection of true ability, even though the two raters do not agree in their ratings. (pp. 53-54)

In this case, raw scores may be underestimated or overestimated if raters of varying severity rate students of the same ability (Engelhard, 1994).

Consistency estimates of inter-rater reliability are based on the assumption that as long as raters are consistent in using the scale by their own definition they do not have to necessarily share a common meaning of the rating. Inter-rater reliability coefficients, however, are often misused as a method to quantify rater effects on observed scores (Zegers, 1991). Stemler (2004) states two important disadvantages to using these methods as a tool for evaluating rater quality: (a) judges may differ systematically in not only raw scores, but the use of the rating scale categories; and (b) possible deflated correlation coefficients can occur due to restricted variability in category usage. Wright and Linacre (1989) highlight Stemler's first disadvantage of CTT methodology:

> . . . this counting of steps says nothing about distances between categories, nor does it require that all test items employ the same rating scale. Whenever four category labels share the same ordering, however else they may differ in implied amounts, they can only be represented by exactly the same step counts, even though, after analysis, their calibrations may well differ. (p. 857)

Therefore, CTT estimates of inter- and intra-rater reliability of a rating scale tell us very little about a rating scale's value since the apparent reliability may be due to errors and biases in rater behavior rather than differences in true scores (Wherry, 1952). Application of the MFR-PC Measurement Model to rater-mediated music performance assessment ratings can help improve objectivity in the measurement process by addressing each of these weaknesses found under the CTT paradigm (see Haiyang, 2010; Huang, Guo, Loadman, & Law, 2014 for direct empirical comparison of CTT and the MFR model).

DATA ANALYSIS PROCEDURE

The Multifaceted Rasch Partial Credit (MFR-PC) Measurement Model is a special formulation of the Many-Facet Rasch (MFR) model (Linacre, 1989/1994). Application of the Partial Credit Model (Wright & Masters, 1982) to Linacre's MFR model extends the analysis to free response alternatives for different raters in the same rating scale (Bond & Fox, 2007). In contrast to the original rating scale (RS) formulation of the model, the Partial Credit (PC) formulation allows the distance between rating scale category thresholds to vary across elements of a specified facet, such as raters. In other words, separate parameterization for each rater within the rating scale can be explored. Statistically, the process of freeing each rater from a rating scale grouping and allowing it to define its own partial credit scale allows for two or more ordered categories to be estimated, potentially reducing misfit and providing better model data fit. The drawback, however, is that the addition of these parameters can limit inference due to unstable difficulty estimates. This is particularly true in instances when all provided response categories are not used (Linacre, 2000; Wright, 1998). According to Wright (1998):

> Each item on a survey or questionnaire represents a universe of other similar items that could have been asked. As we think of these other items, do we place them in the rating scale cluster? Do we impute a particular item's partial credit scale to them? Or do we imagine each of these other possible items to have their own partial credit scales? We are at a loss. But if the original items are modeled to share a rating scale, then we feel secure in imputing that same scale to similar unasked items. (p. 642)

The Partial Credit (PC) model provides the opportunity for a measurement tool to provide different numbers of response opportunities for each item (Bond & Fox, 2007).

Because it is within the family of Rasch measurement models (Wright & Mok, 2004), the MFR-PC model provides a method for examining five required measurement characteristics for rater-invariant measurement including: (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) 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). Adherence to these requirements (i.e., model-data fit) is necessary in order to realize the benefits of the MFR-PC model. In the context of the current study, evidence of model-data fit is necessary in order to interpret differences between the severity of individual raters (research question 1), the difficulty of individual items (research question 2), and the locations of rating scale categories across different raters (research question 3). When good model-data fit is obtained, the MFR-PC model yields invariant measurement for the rater-mediated assessment (Engelhard, 1994), such that these differences can be explored and interpreted.

In particular, a major advantage of the Partial Credit (PC) model is that it can be used to identify differences in the application of the rating scale across different raters. The PC model is specified as follows:

$$\ln\left[\frac{p_{nijmk}}{p_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \gamma_m - \tau_{ik} \qquad (1)$$

where

$\ln[\mathrm{P}_{\mathrm{nijmk}}/\mathrm{P}_{\mathrm{nijmk\text{-}1}}]$ = 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_{ik} = the location on the logit scale where rating scale categories *k* and *k*-1 are equally probable for Rater *i*.

This formulation of the MFR-PC model makes it possible to explore the hypothesis of equidistant rating scale categories across a group of raters. In other words, the PC model provides an empirical test of the hypothesis that a group of raters share a common interpretation of the distance between rating scale categories.

METHOD

The Jazz Big Band Performance Rating Scale (JBBPRS) (Wesolowski, 2016) served as the measurement apparatus for this study (See Figure 1). The JBBPRS consists of 22 items developed from a factor-analytic method of scale construction. Each item was paired with a four-point Likert scale. Responses included "strongly agree," "agree," "disagree," and "strongly disagree."

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. Recordings were carefully screened for audible clarity by the author and two outside evaluators. Recordings consisted of full performances of jazz big bands performing music in a medium swing style.

Experienced raters ($N = 23$) were solicited based upon performance and teaching experience within the jazz idiom. Each rater was the director of a collegiate-level jazz big band had at least 15 years of professional adjudication experience. Raters were supplied with four anonymous recordings (middle school, $n = 1$; high school, $n = 1$, collegiate, $n = 1$; professional, $n = 1$) and asked to evaluate each recording using the JBBPRS to the best of their ability. Recordings were distributed based upon a judging plan recommended by Wright and Stone (1979) and Linacre and Wright (2004) such that the raters formed an incomplete assessment network with links between raters (Engelhard, 1997).

RESULTS

Summary statistics. Table 1 provides the PC-MFR measurement model summary statistics from *FACETS* (Linacre, 2014) for ensembles (θ), raters (λ), items (δ), and school level (γ). The analysis indicated overall significant differences for ensemble, $\chi^2{}_{(22)} = 107.20, p < .05$, raters, $\chi^2{}_{(23)} = 141.10, p < .05$, items, $\chi^2{}_{(21)} = 215.00, p < .05$, and school level, $\chi^2{}_{(3)} = 2115.30$, $p < .05$. The reliability of separation statistics for raters, items, and school level represents the spread of elements within each facet. In the context of general linear modeling, this is comparable to demonstrating a significant main effect where raters, items, and school levels represent independent variables. Moderate to high reliabilities of separation between raters ($REL_{\mathrm{Raters}} = .84$), items ($R\mathrm{EL}_{\mathrm{Items}} = .90$), and school levels ($REL_{\mathrm{School\ Level}} = .99$) indicated that the Jazz Big Band Performance Rating Scale was able to reasonably separate each facet on the underlying latent trait of jazz big band performance achievement. The reliability of separation statistics for ensembles ($REL_{\mathrm{Ensemble}} = .81$) is equivalent to Cronbach's coefficient alpha. Good model data fit is demonstrated by reasonable item mean-square (MSE) ranges for infit and outfit (centering on expected values of 1.0 with a range of .04 to 1.2 for

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, 2015).

TABLE 1. *Summary Statistics from the PC-MFR Model*

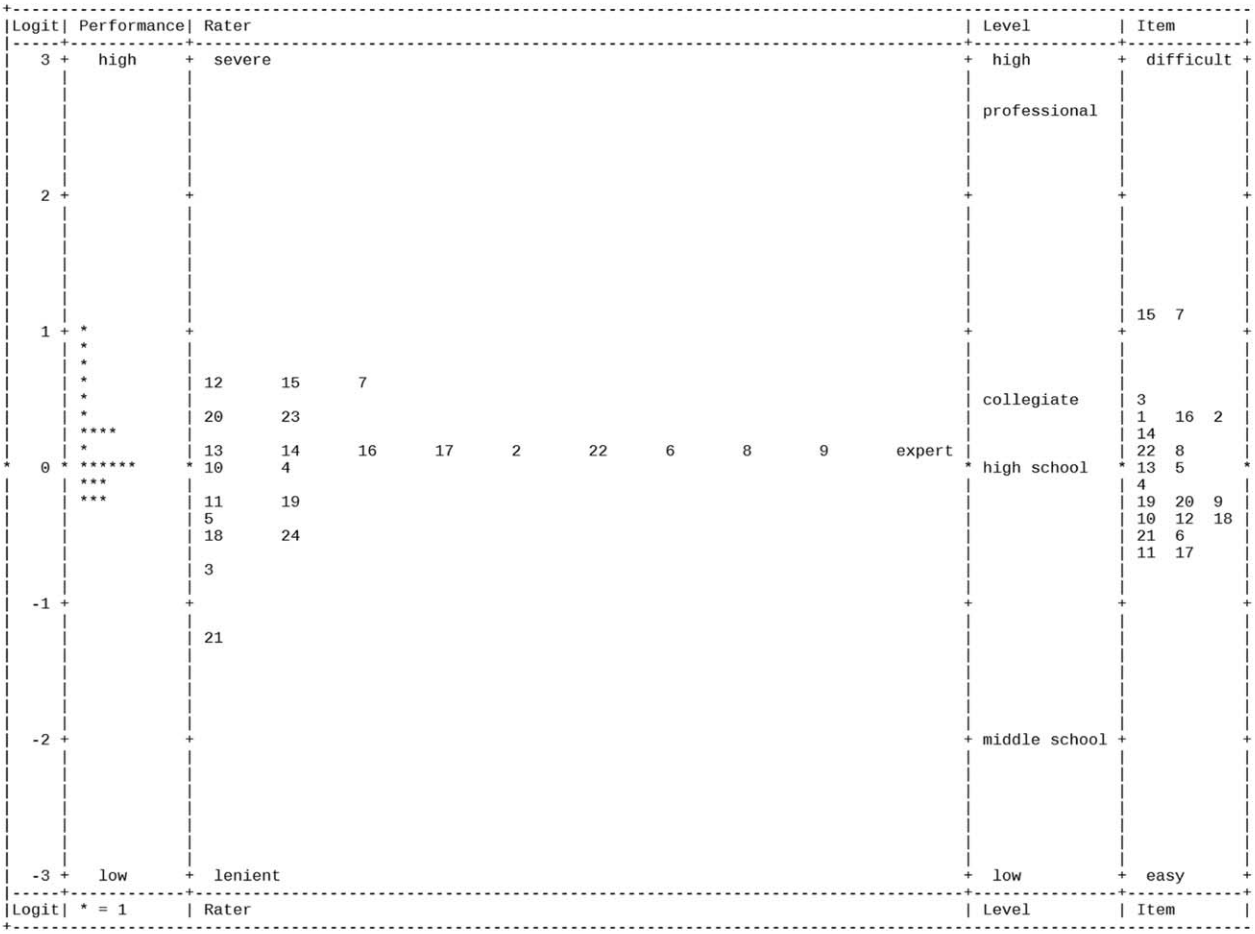
	Facets			
	Ensemble (θ)	Rater (λ)	Item (δ)	School Level (γ)
Measure (Logits)				
Mean	0.17	0.00	0.00	0.27
SD	0.35	0.43	0.48	1.63
N	23	24	22	4
Infit *MSE*				
Mean	0.98	0.99	0.97	0.97
SD	0.19	0.19	0.19	0.08
Std. Infit *MSE*				
Mean	−0.20	−0.10	0.98	0.98
SD	1.40	1.30	0.20	0.07
Outfit *MSE*				
Mean	0.98	1.00	−0.30	−0.60
SD	0.20	0.20	1.50	1.40
Std. Outfit *MSE*				
Mean	−0.20	−0.10	−0.20	−0.30
SD	1.40	1.40	1.50	1.20
Separation Statistics				
Reliability of Separation	.81	.84	.90	.99
Chi-Square	107.2*	147.1*	215.0*	2,115.3*
Degrees of Freedom	22	23	21	3

*$p < .05$

judged assessments) (Wright & Linacre, 1994). Although Outfit MSE for item is just below the acceptable range and considered less productive for measurement (-0.30), Linacre & Wright indicate that is not degrading. However, caution should be taken as it may produce misleadingly high reliability and separation coefficients. Standardized fit statistics (Z*std*) are *t*-tests (reported as *z*-scores) that test the hypothesis of perfect model data fit for predictability of data. Less than the expected score of 0.00 indicates predictability and above 0.00 indicates lack of predictability. All data fits in the range of -1.9 to 1.9 indicating reasonable predictability and good model data fit (Wright & Linacre, 1994).

CALIBRATION OF FACETS

Figure 2 is the variable map for the PC-MFR model generated from the FACETS analysis. The variable map summarizes the results from the PC-MFR model analysis in that it displays the calibrations of each ensemble performance and calibration of all elements (i.e., individual raters, each school level, and each item on the rating scale) within each facet of interest (e.g., raters, school levels, and items) on a log-odds linear scale that represents the latent construct. The log odds metric contains equally spaced units representing the unidimensional latent construct of interest. In this study, the



```
+---------------------------------------------------------------------------------------------------------------------------------
|Logit| Performance| Rater                                                                          | Level         | Item       |
|-----+------------+--------------------------------------------------------------------------------+---------------+------------+
|   3 +   high     +  severe                                                                        +  high         +  difficult +
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                | professional  |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|   2 +            +                                                                                +               +            +
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               | 15  7      |
|   1 + *          +                                                                                +               +            +
|     | *          |                                                                                |               |            |
|     | *          |                                                                                |               |            |
|     | *          | 12      15      7                                                              |               |            |
|     | *          |                                                                                | collegiate    | 3          |
|     | *          | 20      23                                                                     |               | 1   16  2  |
|     | ****       |                                                                                |               | 14         |
|     | *          | 13      14      16      17      2       22      6       8       9       expert |               | 22  8      |
*   0 * ******     * 10      4                                                                      * high school   * 13  5      *
|     | ***        |                                                                                |               | 4          |
|     | ***        | 11      19                                                                     |               | 19  20  9  |
|     |            | 5                                                                              |               | 10  12  18 |
|     |            | 18      24                                                                     |               | 21  6      |
|     |            |                                                                                |               | 11  17     |
|     |            | 3                                                                              |               |            |
|     |            |                                                                                |               |            |
|  -1 +            +                                                                                +               +            +
|     |            |                                                                                |               |            |
|     |            | 21                                                                             |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|  -2 +            +                                                                                + middle school +            +
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|     |            |                                                                                |               |            |
|  -3 +   low      +  lenient                                                                       +  low          +  easy      +
|-----+------------+--------------------------------------------------------------------------------+---------------+------------+
|Logit| * = 1      | Rater                                                                          | Level         | Item       |
+---------------------------------------------------------------------------------------------------------------------------------
```

FIGURE 2. Variable Map for the PC-MFR Model.

latent construct is defined as jazz big band performance achievement. Column 1 (far left) of the variable map shows the logit scale, which serves as the "ruler" onto which all of the facet locations are mapped. Each of the facets displayed in Figure 2 except for the performance facet was anchored at 0.00 logits in order to provide a frame of reference related to other facets in the analysis. The directionality of each facet is indicated at the top of the variable map. The lower end of the log odds scale indicates less of the latent construct, and the higher end of the log odds scale indicates more of latent construct.

Ensembles. The second column of the variable map indicates the spread of performance achievement where each asterisk indicates an individual ensemble performance. Higher logits indicate more of the latent construct (i.e., higher performance achievement) and lower logits indicate less of the latent construct (i.e., lower performance achievement). When data fit the model, the calibrations on the logit scale provide an fair and objective rank ordering of ensembles based upon probabilistic transformation of raw scores on a linear, equal-interval log odds metric. As indicated in Table 2, performance achievement ranged from 1.02 (ensemble 12) to -0.28 (ensemble 23) logits ($M = 0.17$, $SD = 0.25$, $N = 23$). Elements within each facet are considered to be misfit if their infit and/or outfit MSE statistics fall outside of the range of .08 to 1.2 (or standardized infit and outfit MSE of ± 2.00) (Engelhard, 2013). Misfit ensemble performances included performances 4, 7, 11, 15, 17, and 22.

Raters. The third column of the variable map specifies an objective rank ordering of raters based upon severity. The directionality was set opposite of performances (negative), whereby higher logits indicate more severity. As seen in Table 3, rater severity ranged from 0.65 (rater 7, most severe) to -1.19 (rater 21, most lenient). Raters

TABLE 2. *Calibration of the Ensemble Facet*

Ensemble Number	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
12	3.16	1.02	.17	0.94	−0.50	0.99	−0.08
4	2.67	0.86	.16	1.46	2.94	1.54	3.26
17	2.54	0.79	.16	1.21	1.54	1.30	2.01
9	2.05	0.59	.16	0.93	−0.49	0.93	−0.45
5	3.05	0.46	.17	1.02	0.19	1.04	0.37
22	2.85	0.32	.16	1.34	2.27	1.31	2.14
2	2.31	0.31	.15	0.92	−0.64	0.91	−0.70
11	2.85	0.22	.16	0.66	−2.76	0.63	−2.70
19	2.47	0.21	.15	0.82	−1.42	0.81	−1.51
10	2.34	0.20	.15	0.96	−0.27	0.98	−0.12
1	2.55	0.09	.16	0.98	−0.10	0.97	−0.18
16	2.53	0.06	.16	0.87	−0.93	0.91	−0.63
8	2.48	0.04	.14	0.83	−1.33	0.82	−1.46
18	2.82	0.04	.16	0.86	−1.20	0.85	−1.33
6	2.63	0.02	.15	0.79	−1.61	0.80	−1.40
13	2.39	0.00	.16	1.04	0.32	1.08	0.66
20	2.89	−0.05	.16	1.11	0.88	1.05	0.45
7	2.66	−0.11	.16	0.73	−2.23	0.75	−2.04
15	2.35	−0.12	.15	1.21	1.49	1.20	1.37
21	2.70	−0.12	.16	1.00	0.04	0.93	−0.45
14	2.58	−0.26	.18	1.01	0.13	1.06	0.43
3	2.58	−0.27	.15	0.82	−1.47	0.81	−1.57
23	2.48	−0.28	.16	0.98	−0.12	0.98	−0.14
Mean	2.61	0.17	.16	0.98	−0.23	0.98	−0.18
SD	0.25	0.36	.01	0.19	1.39	0.20	1.43

Note. The ensembles are arranged in Measure (achievement) order, from high to low.

11, 18, and 19 demonstrated muted rating patterns as evidenced by Infit MSE less than .80. Raters 2, 3, 24, and 10 demonstrated haphazard use of the rating scale as indicated by Infit MSE greater than 2.0.

School level. The fourth column of the variable map presents calibrations of the school level facet. As seen in Table 5, school level placement along the log odd latent variable makes intuitive sense, demonstrating a logical ordering of least to most performance achievement: middle school (−2.02), high school (0.04), collegiate (2.77), professional (3.47). The mean achievement value for all performances was 2.64 (*SD* = 0.70).

Items. The fifth column of the variable map indicates the spread of item difficulty. As seen in Table 4, the calibrations of items range from 1.12 (item 15, most difficult) to −0.60 (item 11, easiest). Items were anchored at 0.00 in order to provide a clear reference point. The placement of items on the variable map is directly related to the measurement latent construct of interest (i.e, performance achievement). Therefore, it can be interpreted that as the logit values increase on each item, the level of ensembles' performance achievement increases on each item.

RATING SCALE STRUCTURE BY RATER

Figure 3 presents the structure of rating scale usage for each of the 24 raters. The unique category coefficient thresholds for each rater are indicated by a tao (τ). Each τ is calibrated to the logit scale. τ_1 represents the threshold between a 1 (*strongly disagree*) and 2 (*disagree*). τ_2 represents the threshold between a 2 (*disagree*) and 3 (*agree*). τ_3 represents the threshold between a 3 (*agree*) and 4 (*strongly agree*). By reviewing these structures as well as frequency data, average observed and expected measures, and outfit MSE as provided in Table 6, one can evaluate specific rater behaviors and infer overall rater quality with valid, reliable, and linear empirical evidence. As an example, rater 4 and rater 6 both demonstrate a wide usage of category 3. In this case, if an ensemble was evaluated only by raters 4 and 6 they may receive a score of 3; however, if the same ensemble was evaluated by rater 5, they would receive a score of 2. Category probability curves plots can also aid in evaluating category usage of raters (see Figure 4). Specifically, each of the plots in Figure 4 shows the probability for a given rater to assign a rating in categories 1 - 4 (*y*-axis), given student locations on the latent variable (*x*-axis), with each category illustrated as a separate curve. For

TABLE 3. *Calibration of the Rater Facet*

Rater Number	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
7	2.42	0.65	.17	0.88	−0.87	0.87	−0.92
12	2.44	0.62	.17	0.93	−0.44	0.94	−0.38
15	2.41	0.57	.18	1.07	0.51	1.06	0.47
23	2.61	0.40	.18	0.90	−0.72	0.91	−0.64
20	2.45	0.34	.18	1.03	0.23	1.02	0.18
6	2.70	0.16	.19	1.12	0.74	1.18	1.04
9	2.58	0.16	.18	0.97	−0.15	0.91	−0.53
13	2.68	0.15	.18	0.96	−0.19	0.90	−0.50
22	2.67	0.15	.17	0.99	−0.05	1.00	0.03
C	2.60	0.14	.07	0.93	−1.26	0.92	−1.38
8	2.66	0.14	.16	0.86	−0.90	0.97	−0.15
16	2.58	0.11	.17	0.62	−2.98	0.63	−2.90
2	2.56	0.09	.19	1.24	1.67	1.24	1.69
14	2.56	0.09	.18	1.05	0.39	1.09	0.66
17	2.56	0.09	.18	0.89	−0.79	0.90	−0.69
10	2.62	0.02	.17	1.40	2.45	1.38	2.30
4	2.70	−0.05	.20	1.17	1.17	1.18	1.19
11	2.70	−0.21	.18	0.70	−2.19	0.68	−2.19
19	2.51	−0.30	.18	0.73	−1.98	0.72	−2.03
5	2.86	−0.42	.17	0.88	−0.81	0.81	−1.15
18	2.65	−0.45	.19	0.77	−1.70	0.79	−1.51
24	2.86	−0.49	.20	1.25	1.56	1.28	1.74
3	2.27	−0.80	.18	1.25	1.71	1.36	2.08
21	2.89	−1.19	.17	1.13	1.00	1.18	1.26
Mean	2.61	0.00	.17	0.99	−0.15	1.00	−0.10
SD	0.15	0.44	.02	0.19	1.36	0.21	1.40

Note. The raters are arranged in Measure (severity) orer, from high to low.

TABLE 4. *Calibration of the School Level Facet*

School Level	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
Middle School	1.78	−2.02	.06	0.96	−0.74	0.98	−0.39
High School	2.55	0.04	.06	1.08	1.43	1.08	1.46
Collegiate	2.77	0.50	.06	0.86	−2.57	0.89	−1.89
Professional	3.47	2.57	.08	0.98	−0.43	0.98	−0.28
Mean	2.64	0.27	.07	0.97	−0.58	0.98	−0.28
SD	0.70	1.88	.01	0.09	1.64	0.08	1.37

Note. The school levels are arranged in Measure (achievement) order, from high to low.

example, the plot for the Common rater shows that students who are lower on the logit scale are most likely to receive a rating in category 1, whereas the probability for a rating in this category decreases as student locations increase. As an example, Figure 4 provides five examples of diverse rating scale use. The common rater demonstrates a consistent interpretation and application of the rating scale. Rater 3 demonstrates no use of category 4. Rater 8 demonstrates limited use of category 2. Rater 11 demonstrates limited use of category 3. Rater 24 demonstrates a central tendency by overuse of categories 2 and 3. The inconsistency of raters' use of rating scale categories as illustrated by erratic category probability curves indicates that ensembles with varying levels of performance achievement (in adjacent rating scale categories) may not always be distinguished. Conversely, ensembles of the same achievement level may be inappropriately distinguished. The lack of observations in the extreme categories might be a contributing factor to misfit. However, misfit is more directly related to unexpected observations, where the rater's observed rating doesn't match what would be expected by the model for that student, given their achievement level.

Quality control of rating scale structure was further verified with the demonstration of ascending mean observed scores corresponding to each rating category:

TABLE 5. *Calibration of the Item Facet*

Item Number	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
15	2.19	1.12	.15	1.11	0.91	1.14	1.07
7	2.20	1.09	.15	1.08	0.66	1.00	0.02
3	2.43	0.47	.15	1.34	2.46	1.30	2.22
16	2.45	0.42	.15	0.75	−2.10	0.79	−1.73
1	2.47	0.37	.15	1.05	0.40	1.05	0.43
2	2.48	0.35	.15	0.84	−1.30	0.85	−1.19
14	2.50	0.28	.15	1.06	0.51	1.12	0.97
8	2.56	0.14	.15	0.66	−2.99	0.69	−2.66
22	2.58	0.07	.15	0.75	−2.13	0.78	−1.86
5	2.60	0.02	.15	0.86	−1.14	0.86	−1.11
13	2.63	−0.05	.15	1.15	1.16	1.26	1.93
4	2.67	−0.17	.15	0.93	−0.54	0.97	−0.23
9	2.68	−0.19	.15	1.16	1.23	1.16	1.18
20	2.69	−0.22	.15	0.97	−0.16	1.00	0.01
19	2.70	−0.24	.15	0.73	−2.25	0.77	−1.84
18	2.75	−0.38	.16	0.91	−0.67	0.90	−0.75
10	2.76	−0.41	.16	0.84	−1.23	0.85	−1.13
12	2.77	−0.43	.16	1.10	0.81	1.14	1.02
6	2.79	−0.50	.16	0.95	−0.31	0.92	−0.58
21	2.81	−0.55	.16	0.74	−2.19	0.75	−2.01
17	2.82	−0.58	.16	1.43	2.94	1.47	3.06
11	2.83	−0.60	.16	0.94	−0.43	0.90	−0.73
Mean	2.61	0.00	.15	0.97	−0.29	0.99	−0.18
SD	0.18	0.49	.00	0.20	1.56	0.20	1.52

Note. The items are arranged in Measure (difficulty) order, from high to low.

strongly disagree ($n = 52$; $M = -1.60$), agree ($n = 178$; $M =$ -.65), agree ($n = 195$; $M = .86$), and strongly agree ($n = 81$; $M = 2.10$).

Discussion

Measurement in the behavioral sciences consists solely of latent variables. Unlike measurement in the physical sciences, there can never be a comparison of directly observable qualities with known standards. Assuming these latent qualities truly exist is common in the psychological sciences as evidenced by researchers too often mistaking raw scores for linear measures. However, a latent variable is not to be mistaken as an empirical quality with a hypothesis. The only way to define a latent construct is through a model. In particular, measurement models can formulate hypothetical relations between observed data and the latent construct.

In the context of rater-mediated music performance evaluations, it is desirable for the raters to evaluate performances with consistency and objectivity in scoring in order to most accurately reflect the true ability of the objects of measurement (e.g., music performances). However, the problem with rater-mediated music performance assessments is that raters' observed scores are often associated with characteristics of the raters themselves and not necessarily with the true scores of the evaluated performances (Engelhard, 2002). The use of raters' observed scores as a means to define a latent construct and evaluate music ensembles assumes an overall acceptance that raters are using rating scales in a comparable manner and observed scores are equivalent to true scores. As evidenced in this paper, the data demonstrates that rating scale structures can vary by rater and each rater demonstrates unique tendencies of leniency/severity. With a partial credit model, the evaluation of raters' use of rating scale structure is underscored by the concept of parameterization. In particular, the use of the partial credit model not only addresses stability estimates of raters' leniency/severity, but also provides a method for exploring construct validity of the meaningful difference in rating scale structure. The original response options provided to raters included a 4-point Likert scale format: *strongly disagree, disagree, agree,* and *strongly agree*. The use of the partial credit model not only allowed for the transformation of the categorical (i.e., ordinal) observed responses to continuous (i.e., interval) measures, but also provided an additional parameter in order to provide evidence of raters' unique difficulty estimates on each item of the

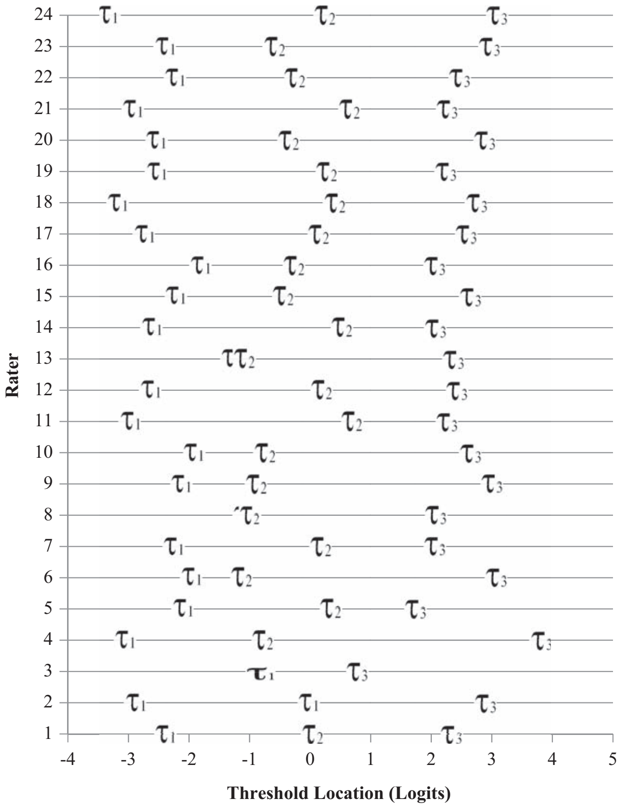


FIGURE 3. Threshold locations for the Rater Facet. *Note:* No τ_3 estimate was calculated for Rater 3, because this rater did not assign ratings in Category 4 (Rating = 4).

scale. Substantively, this is important because it: (a) provides empirical evidence that raters, regardless of "expert" status, do not share the same interpretation of rating scale structure; (b) provides a more precise estimation of true performance providing a more clearly defined and secure rating scale; and (c) has an significant effect on the hierarchical ordering of the severity of raters. Additionally, this technique provides evidence that traditional estimates of rater consensus do not provide substantive meaning regarding the precision of true score performance estimates that can be used to inform rater training, rater monitoring, and the interpretation of rater-assigned scores.

Discrepancies resulting from raters' usage and application of a rating scale can bring about serious ramifications pertaining to fairness. On these occasions, ensembles may be receiving diagnostic data that is based more on the quality of the rater rather than their true performance ability. In high stakes music performance assessments, deserving ensembles and/or persons subject to severe raters may not receive the supporting rating criteria for successful passing. Oppositely, non-deserving ensembles subject to raters demonstrating lenient tendencies may receive over-inflated scores. Quality control and evaluation of raters is therefore necessary for fair, accurate, and precise evaluations. The Rasch model has high standards for misfit due to the five requirements for invariance. "Better model fit" may result using other polytomous item response measurement models. However, these models do not use the same set of strict requirements as the MFR-PC measurement model. The strength of the invariant approach to measurement using Rasch models is that raters, items, and performances are flagged as misfit and and can therefore be diagnosed with more analysis.

Other disciplines managing rater-mediated, high stakes standardized assessments (e.g., national writing, science, and medical assessments), raters continually evaluate performances while concurrently enduring ongoing training and quality control evaluation. Specifically in the field of writing assessment, use of anchors sets of essays as exemplars to clearly demonstrate degrees of proficiency for each item have been shown to improve rater accuracy and uniformity within the application of the rating scale (e.g., Johnson, Penny, & Gordon, 2009; Osborn Popp, Ryan, & Thompson, 2009). It is suggested that in order to improve objectivity in music performance assessment, a system of benchmark anchor recordings along with written commentaries that demonstrate proficiency levels for each item included in the rating scale be developed in order to train and manage the quality of raters. Annotated commentaries can provide a specific rationale for the specific examples. The development of annotated commentaries and sample anchor recordings may serve multiple purposes beyond concrete examples for the improvement of objectivity in measurement and understanding of scoring criteria: (a) to provide clear levels of proficiency for teachers, students, and evaluators; (b) to inform instruction and instructional planning; (c) to provide common benchmarks of student achievement that may promote student ensemble growth and development; (d) to train pre-service teachers in music teacher preparation programs; (e) to garner a more clear understanding of expectations for achievement at multiple grade and and/or ensemble levels; and (f) to promote professional dialogue regarding the improvement of benchmarks, standards, and expectations.

In experimental research that relies on the use of raters, whether it be for gathering evaluative or attributional data, overly severe and lenient raters under CTT data analysis methods may provide critical type I or type II errors that can occur undetected. Analysis and understanding of the varying interpretations of rating scale

TABLE 6. *Rater Behavior of Category Usage, Average Observed and Expected measures, and Outfit MSE*

	Category Usage (%)				Average Observed Measure (Average Expected Measure)				Outfit *MSE*			
Rater	1	2	3	4	1	2	3	4	1	2	3	4
C	52 (10)	178 (35)	195 (39)	81 (16)	−1.60 (−1.66)	−.65 (−.56)	.86 (.83)	2.10 (2.00)	1.00	.80	.90	.90
1	9 (10)	31 (35)	38 (43)	10 (11)	−1.71 (−2.29)	−.72 (−.75)	.76 (.95)	2.18 (2.08)	1.60	1.20	1.30	1.00
2	6 (7)	25 (28)	46 (52)	11 (13)	−2.05 (−2.08)	−.74 (−1.02)	.78 (.92)	3.41 (3.46)	.90	1.60	1.00	1.10
3	16 (18)	32 (36)	40 (45)	–	−.26 (−.62)	.47 (.50)	1.30 (1.42)	–	1.70	1.40	1.10	—
4	6 (7)	27 (31)	28 (32)	27 (31)	−.59 (−.96)	−.31 (−.16)	.84 (.97)	2.68 (2.47)	1.30	.70	.70	.70
5	12 (14)	15 (15)	48 (55)	13 (15)	−2.14 (−2.25)	−1.13 (−.97)	1.14 (.97)	1.92 (2.46)	1.60	1.00	.90	1.40
6	12 (14)	37 (42)	29 (33)	10 (11)	−1.58 (−1.58)	−.76 (−.62)	.70 (.50)	1.39 (1.43)	1.00	.80	.80	.90
7	17 (19)	15 (17)	37 (42)	19 (22)	−.207 (−1.68)	−.35 (−.69)	.85 (.58)	1.77 (2.22)	.40	1.30	.90	1.40
8	12 (14)	22 (25)	45 (51)	9 (10)	−2.20 (−2.01)	−.79 (−.91)	.65 (.62)	1.85 (2.03)	.70	1.20	.70	1.00
9	12 (14)	21 (24)	43 (49)	12 (14)	−1.43 (−1.85)	−.92 (−.75)	.95 (.72)	.98 (1.94)	1.80	1.00	1.10	1.70
10	4 (5)	37 (42)	28 (32)	19 (22)	−1.27 (−1.15)	−.23 (−.23)	.77	3.11 (2.55)	.90	.80	.70	.40
11	10 (11)	38 (43)	31 (35)	9 (10)	−1.79 (−1.80)	−.76 (−.68)	.75 (.65)	1.62 (1.64)	1.00	.80	1.00	1.00
12	17 (19)	14 (16)	37 (42)	20 (23)	−2.18 (−2.06)	−.89 (−.81)	.89 (.63)	2.56 (2.89)	.70	1.00	.70	1.30
13	10 (11)	36 (41)	25 (28)	17 (19)	−1.60 (−1.91)	−.45 (−.62)	.25 (.94)	2.82 (2.34)	1.20	1.00	1.80	.40
14	18 (20)	27 (31)	32 (36)	11 (13)	−2.49 (−2.46)	−.91 (−1.08)	.50 (.64)	2.34 (2.29)	.90	1.30	1.10	.90
15	15 (17)	23 (26)	34 (39)	16 (18)	−2.31 (−2.00)	−.73 (−.62)	.90 (.78)	1.99 (1.78)	.60	.60	.60	.80
16	8 (9)	35 (40)	33 (38)	12 (14)	−2.31 (−1.75)	−.38 (−.64)	.55 (.85)	2.60 (2.16)	.60	1.00	1.20	.60
17	4 (5)	37 (42)	33 (38)	14 (16)	−1.46 (−1.40)	−.49 (−.40)	1.03 (1.07)	2.96 (2.59)	1.00	.80	.90	.60
18	11 (13)	37 (42)	24 (27)	16 (18)	−1.53 (−1.59)	−1.08 (−.93)	.79 (.85)	2.97 (2.56)	1.10	.50	.70	.50
19	7 (8)	37 (42)	41 (47)	3 (3)	−.98 (−1.18)	−.61 (−.56)	.05 (.03)	.43 (.47)	1.10	1.00	1.00	1.00
20	3 (3)	27 (31)	35 (40)	23 (26)	−1.00 (−1.21)	.29 (.01)	1.30 (1.59)	2.08 (2.01)	1.10	1.10	1.90	.90
21	7 (8)	28 (32)	40 (45)	13 (15)	−1.27 (−1.21)	−.32 (−.42)	.58 (.70)	2.12 (1.93)	1.00	1.30	.80	.90
22	11 (13)	26 (30)	37 (42)	14 (16)	−1.95 (−1.94)	−1.20 (−1.01)	.96 (.78)	2.81 (2.92)	.90	.70	.90	1.10
23	3 (3)	27 (31)	37 (42)	21 (24)	−1.77 (−1.59)	−.30 (−.46)	1.67 (1.58)	3.00 (3.35)	.90	1.50	1.30	1.20

Note. Category 1 = *strongly disagree*; Category 2 = *disagree*; Category 3 = *agree*; Category 4 = *strongly agree*

usage can provide more informed prescriptions for the improvement of learning, teaching, and human understanding. Additionally, the evaluation and interpretation of fit indices and standardized scores can aid in detecting a wide range of rater effects in addition to rater severity. These can include rater accuracy, halo effects, central tendencies, and restrictions of range. These indices of rater effects can be used to better inform results of experimental studies.

A benefit of the using Rasch measurement theory as an approach to develop measurement tools under such premises is the ability to include new items that can be validated with additional testing. This allows for the refinement of a measure as well as the ability to design testlets appropriate for various facets of inquiry. Methodologically, the MFR-PC model can be used to evaluate local dependencies between items within a testlet. Furthermore, when using rater-mediated assessment data to develop and refine measures, arbitrary dependencies can be evaluated with raters. Therefore, as an extension to this study, it is recommended that differential rater functioning be evaluated. Investigation into differential rater functioning (i.e., differential leniency/severity of raters within various subgroups such as school level, gender, instrument, etc.) may reveal systematic subpatterns. Such analysis may provide more detailed information on the behavior of both items and raters, thereby further improving objectivity in rater-mediated music assessment processes. Furthermore, following the methodology of Eckes (2008), an application of a cluster analysis to the calibrated data may provide insight into rater cognition and various structures that emerge of how groups of raters respond to performances. Constructing common categorical representations of groups of raters based upon scoring tendencies may provide important information into how groups of raters can be calibrated based upon disposition. In instances when time or operational costs do not allow for strict rater training, categorical representations of raters may provide a foundation for a succinct and expeditious method for recalibrating raters.

Although Rasch modeling was first developed in the 1960s (Rasch, 1960) and has received significant attention and application in the social, behavioral, and health sciences, it has received little attention in music research. The MFR-PC measurement model provides a sound theoretical basis for examining rater behaviors and

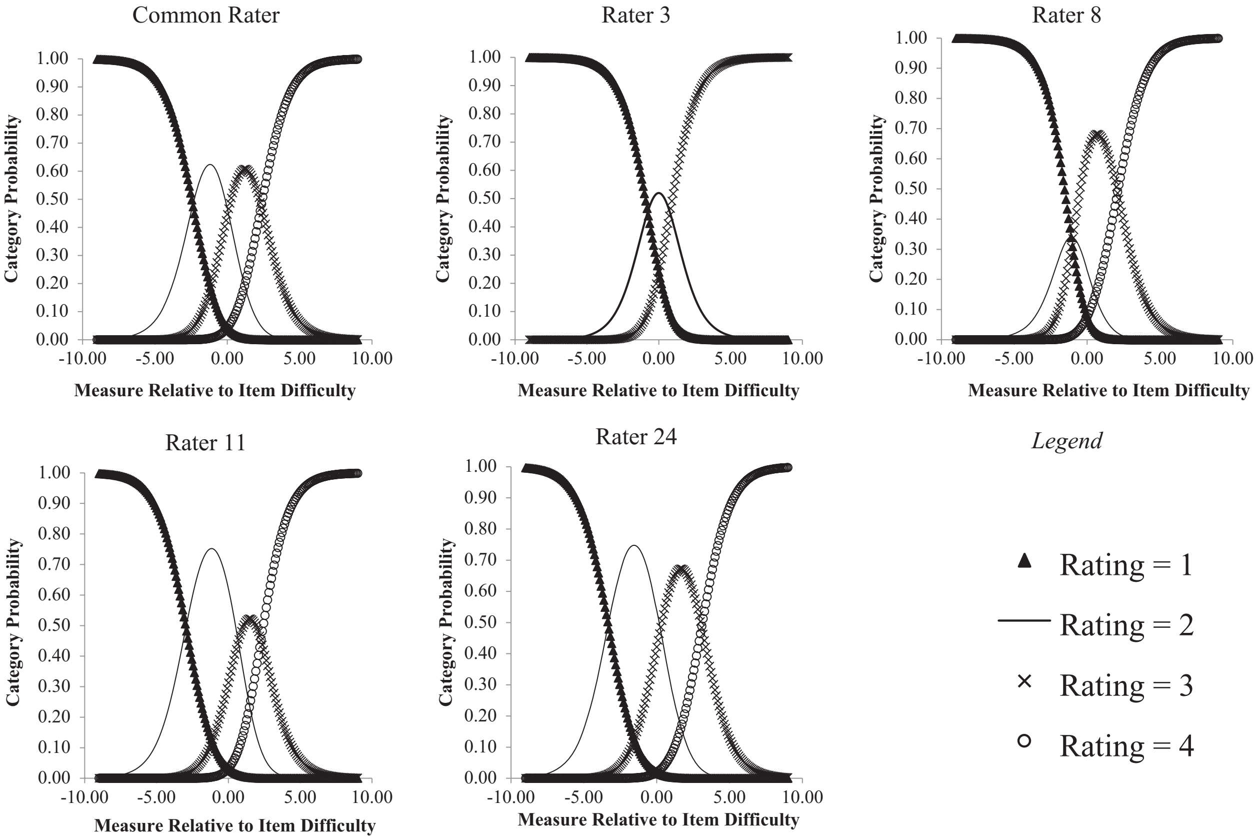


FIGURE 4. Sample category probability curves

judgments. In order to develop clear measures for latent construct measurement in music, all facets of interest in the measurement model must be invariant across raters; otherwise, construct-irrelevant variance can influence the observed data and go undetected. Invariant measurement is supported only when data fit the requirements of the model. Underlying the principle of rater-invariant measurement of performances is the notion of fundamental measurement. Fundamental measurement implies that the collected data should fit the model in order to achieve properties of invariance, not that the model should be manipulated to fit the collected data. The requirements of the model (i.e., the measurement ideal) are of upmost importance in order to demonstrate data sufficiency for the specification of the intended measurement. Therefore, use of an ideal-type model such as the MFR-PC measurement model is necessary for implying rater-invariant measurement of musical performances.

Author Note

Correspondence concerning this article should be addressed to Brian C. Wesolowski, University of Georgia, 250 River Road, Athens, GA 30602. E-mail: bwes@uga.edu

References

Aruffo, C., Goldstone, R. L., & Earn, D. J. D. (2014). Absolute judgment of musical interval width. *Music Perception*, *32*, 184-198.

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, *51*(2), 137-150.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Bishop, M. A., & Trout, J. D. (2005). *Epistemology and the psychology of human judgment*. Oxford: Oxford University Press.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco, CA: Holden-Day.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

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

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago, IL: Chicago University Press.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, *21*(1), 22-29.

Campbell, S. K., Kolobe, T. H., Osten, E. T., Lenke, M., & Girolami, G. L. (1995). Construct validity of the Test of Infant Motor Performance. *Physical Therapy*, *75*, 585-596.

Castro, S. L., & Lima, C. F. (2014). Age and musical expertise influence emotion recognition in music. *Music Perception*, *32*, 125-142.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, *28*, 207-222.

Conrad, D. (2003). Judging the judges: Improving rater reliability at music contests. *NFHS Music Association Journal*, *20*(2), 27-31.

Davidson, J. W., & Coimbra, D. D. C. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musicae Scientiae*, *5*, 33-53.

Descartes, R., & Cottingham, J. (1986). *Meditations on first philosophy.* Cambridge, UK: Cambridge University Press.

DiStefano, C., Greer, F. W., Kamphaus, R. W., & Brown, W. H. (2014). Using Rasch rating scale methodology to examine a behavioral screener for preschoolers at risk. *Joural of Early Intervention*, *36*, 192-211.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*, 268-272.

Eckes, T. (2008). *Rater types in writing performance assessments: A classification approach to rater variability. Language Testing*, *25*(2), 155-185.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175-196.

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

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* (pp. 261–287). Mahwah, NJ: Erlbaum.

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

Engelhard, Jr., 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.

Fiske, H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*, 256-263.

Fiske, H. E. (1978). The effect of a training procedure in musical performance evaluation on judge reliability. Brantford, Ontario, Canada: *Ontario Education Research Council Report*

Fiske, H. E. (1979). Musical performance evaluation ability: Toward a model of specificity. *Bulletin of the Council for Research in Music Education*, *59*, 27-31.

Fiske, H. E. (1983). Judging musical performance: Method or madness? *Update: Applications of Research in Music Education, 1*(3), 7-10.

Flores, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth Musical Competition: How fair is the final ranking? *The Statistician*, *45*(1), 97-104.

Forbes, G. W. (1994). Evaluating music festivals and contests - are they fair? *Update: Applications of Research in Music Education*, *12*(2), 16-20.

Freedman, S. W., & Calfee, R. C. (1983). Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York: Longman.

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *Journal of Neuroscience*, *34*(32), 10573-10581.

Freud, S. (1920). *Beyond the pleasure principle*. New York: Liveright.

Ginsburgh, V., & Weyers, S. (2007). Quantitative approaches to valudation in the arts, with an application to movies. In M. Hutter & D. Throsby (Eds.), *Beyond price: Value in cluture, economics, and the arts* (pp. 179-199). Cambridge, UK: Cambridge University Press.

Guliford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Haiyang, S. (2010). An application of Classical Test Theory and Many-Facet Rasch Measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics (Bimonthly)*, *33*(2), 87-102.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41*(3), 337-373.

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.

Haswell, R. (2001). *Beyond outcomes: Assessment and instruction within a university writing program* (R. Haswell, Ed.). Westport: CT: Ablex.

Henning, G. (1997). Accounting for nonsystematic error in performance ratings. *Language Testing*, *13*(1), 53-63.

Hogarth, R. (1987). *Judgment and Choice* (2nd ed.). New York: John Wiley and Sons.

Huang, T., Guo, G., Loadman, W., & Law, F. (2014). Rating score data analysis by Classical Test Theory and Many-Facet Rasch Model. *Psychology Research*, *4*(3), 222-231.

Hutchins, S., Hutka, S., & Moreno, S. (2014). Symbolic and motor contributions to vocal imitation in absolute pitch. *Music Perception*, *4*, 695-701.

Jackson, R. S. (2009). *Wine tasting: A professional handbook*. Burlington, MA: Academic Press.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks*. New York: Guliford Press.

Kant, I., & Meredith, J. C. (1986). *The critique of judgement*. New York: Clarendon Press.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404-426.

King, S. E., & Burnsed, V. (2007). 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*, 27-33.

Labbe, C., & Grandjean, D. (2014). Musical emotions predicted by feelings of entrainment. *Music Perception*, *32*, 170-185.

Latimer, M. E., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, *58*(2), 168-183.

Linacre, J. M. (1989/1994). *Many facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2000). Comparing "partial credit models" (PCM) and "rating scale models" (RSM). *Rasch Measurement Transactions*, *14*(3), 768.

Linacre, J. M. (2002). Judge ratings with forced agreement. *Rasch Measurement Transactions*, *16*(1), 857-858.

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

Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many-facet data. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theories, models, and applications* (pp. 296-321). Maple Grove, MN: JAM Press.

Looney, M. A. (1997). Objective measurement of figure skating performance. *Journal of Outcome Measurement*, *1*(2), 143-163.

McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical excellence: Strategies and techniques to enhance performance* (pp. 61-82). Oxford, UK: Oxford University Press.

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education*, *10*(1), 12-24.

Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, *17*(2), 173-181.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

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.

O'Neill, P. (2002). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1*(1), 47-65.

Osborn Popp, S. E., Ryan, J. M., & Thompson, M. S. (2009). The critical role of anchor paper selection in writing assessment. *Applied Measurement in Education*, *22*(3), 255-271.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, *30*, 71-83.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

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.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18*(1), 46-56.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research Evaluation*, *9*(4), 1-19.

Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York: John Wiley and Sons.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, *21*, 21-41.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception*, *25*, 13-29.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions, *Trends in Cognitive Science, 12*, 455-460.

Wesolowski, B. C. (2016). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music*, *44(3)*, 324-339.

Wesolowski, B. C., Wind, S. A., & Engelhard, J. G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 39-47.

Wherry, R. J. (1952). *The control of bias in ratings: VII. A theory of rating* [PRB Final Report No. 922]. Columbus, OH: Ohio State University Research Foundation.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Taylor & Francis Group, LLC.

Wright, B. D. (1998). Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)? *Rasch Measurement Transactions*, *12*(3), 641-642.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857-860.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Zegers, F. E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement*, *15*(4), 321-333.