Bulletin of the Council for Research in Music Education
Spring 2017 No. 212

Evaluating Differential Rater Functioning Over Time in the Context of Solo Music Performance Assessment

Brian C. Wesolowski
University of Georgia
Athens, GA

Stefanie A. Wind
University of Alabama
Tuscaloosa, AL

George Engelhard, Jr.
University of Georgia
Athens, GA

ABSTRACT

Rater variability studies in the context of music performance assessment treat rater effects as static characteristics of raters, where the effects occur similarly across each assessed performance. The purpose of this study was to investigate expert raters' (N = 13) differential severity/leniency as dynamic processes, where the rater effects occur over time. In particular, we sought to examine the manifestation of group and individual variability using a class of rater effects referred to as differential rater functioning over time (DRIFT). DRIFT refers to the changes in rater performance in relation to a parameter of time. Three classes of Multifaceted Rasch (MFR) models were specified in order to explore differences in raters' systematic changes in their interpretation of a 4-point rating scale structure across a 5-day rating session: (a) time-static model, (b) rater-by-time interaction model, and (c) partial credit model for time points. Results indicated a significant difference in severity/leniency across time for both the group of raters as a whole and some individual raters. Overall, raters demonstrated a general trend of decreasing severity over the 5-day rating session. Interaction analyses suggested that differential severity/leniency existed for both the raters as a group and for 9 out of the 13 individual raters. Of the total 65 potential pairwise interaction terms examined between raters and days, 21 (33.31%) were found to be statistically significant. Ten interactions systematically underestimated the performances and 11 interactions systematically overestimated the performances. Implications for the improved fairness of ratings in music assessment contexts are discussed.

In formal music assessment contexts, raters can work between 8 to 16 hours per day for several consecutive days (Barnes & McCashin, 2005). According to Wolfe, Moulder,

and Myford (1999), "When the rating task takes place over the period of several hours or several days, concern may arise about the comparability of ratings both between and within raters over time" (p. 3). Differential rater functioning over time (DRIFT; Wolfe et al., 1999; Wolfe, Myford, Engelhard, & Manalo, 2007) needs to be taken into account in order to defend against the threat of construct irrelevant variability in the context of extended rating sessions. A measurement framework based on invariant measurement is particularly well suited to the empirical investigation of these concerns. Specifically, invariant measurement provides a framework for exploring differential rater functioning over time in terms of three major considerations.

First, a measurement framework based upon invariant measurement allows for the joint calibration of facets that can independently measure each rater's severity/leniency and each performer's achievement (i.e., a direct comparison of performers in a manner that does not depend on which rater happened to evaluate the performance). Second, the framework provides empirical evidence of each rater's unique interaction with the evaluative cues set forth in the measurement instrument that suggests departures from invariant measurement (i.e., an empirical investigation of rater effects through a quantitative definition of each rater's unique behaviors). Third, the framework allows for the interaction of the calibrations of each rater's severity/leniency with points in time to be detected (i.e., a direct comparison of a rater behavior from one time point to another time point). In order to meet the requirements for invariant measurement (discussed further below), raters' scores must remain consistent across performers, across the measurement instrument, and over time.

Similar to other performance assessment contexts, raters presiding over formal music performance assessments are most often solicited from a pool of content experts, as the field of music is primed to expect fair evaluations from those demonstrating success in the field (Conrad, 2003; Fautley, 2010). Characteristics that deem raters an "expert" include but are not limited to years of experience, success as an ensemble director, and the ability to identify, diagnose, and communicate prescribed solutions to common performance problems (Kruth, 1970). Depending on specific state and/or district Music Educator Association protocols or the context of the performance assessment (e.g., audition, jury, recital, competition, etc.), raters are expected to provide evaluative feedback consisting of either qualitative marking schemes (i.e., audio tape recording of real-time commentary of performances accompanied by written narrative), quantitative marking schemes (i.e., use of an empirically based rating form), or a hybrid of both marking schemes. For qualitative marking schemes, the use of content-expert adjudicators is significant as they have the expertise to instantly provide performers and teachers with valuable best-practice diagnostic information that cannot be obtained by empirical evaluations alone. Such comments include praise, encouragement, and personalized strategies aimed at improving specific elements of the musical performance (Ellis, 1997).

Quantitative marking schemes, however, are more often the subject of attention (Ellis, 1997; McPherson & Schubert, 2004; McPherson & Thompson, 1998).

Empirical results have been shown to improve student motivation, increase student self-efficacy, and enhance the quality of student musicianship (Austin, 1988; Banister, 1992; Franklin, 1979; Howard, 1994; Hurst, 1994; Sweeney, 1998). Considerations of empirical results additionally impact repertoire considerations (Crochet, 2006), classroom performance objectives, long-range goals, and curricular reform (Abeles, Hoffer, & Klottman, 1994; Howard, 2002). Beyond the direct effect on teaching and instruction, the empirical results of quantitative marking schemes more broadly influence community and administrator perceptions of teacher effectiveness and program quality (Boyle, 1992; Burnsed, Hinkle, & King, 1985; Kirchhoff, 1988). Furthermore, the results and conditions of performance assessments can have a strong impact on educators' job security (Barnes & McCashin, 2005; Burnsed et al., 1985). Therefore, it is important for empirical marking schemes to represent the true performance as accurately and precisely as possible. In order for this to occur, music assessment contexts using quantitative marking schemes as a means to evaluate student achievement and proficiency should be managed in a manner that provides as valid, reliable, precise, and fair as any high-stakes performance assessment schemes in other academic fields.

Music performance assessments use constructed response (CR) measures as a means to evaluate performers' abilities. Unlike traditional selected-response measures where items can be coded dichotomously as correct and incorrect or coded polytomously as an ordered response, CR music performance assessment measures require rater intermediation. Rater-mediated assessment frameworks can be conceptualized as a lens model, where judgmental precision and accuracy are bound by raters' independent, observed ratings nested within a set of preestablished evaluative cues (Brunswik, 1952; Engelhard, 2013; Hogarth, 1987). Observed ratings are based on raters' value judgments, guided by their unique interpretations of performance proficiency levels and cues prompted by the measurement instrument. Because the cues set forth in the measurement instrument operationally define the latent construct (e.g., music performance achievement), raters' proper interpretation and use of the cues is necessary for supporting validity evidence of the assessment context. However, in instances when multiple raters independently evaluate the same musical performance, rarely will they perfectly agree (Bergee, 1989, 1997, 2003, 2007; Flores & Ginsburgh, 1996; Hash, 2012; Latimer, Bergee, & Cohen, 2010; Norris & Borst, 2007). Rater variability contains both systematic and probabilistic elements; therefore, an understanding of each raters' quantitative characteristics is required for valid assessment practices (Linacre, 1989). As such, the latent construct being measured (e.g., music performance achievement) and the validity of the measure itself can be obscured through unwanted variability in observed scores (Lane & Stone, 2006).

Across performance assessment contexts in general, raters' schemata vary in the use of evaluation cues and the cognitive processes by which the scoring is based, causing fundamental validity concerns with the misconception that observed scores are "measures" (Wolfe, 1997). Under quantitative marking schemes, content-expert raters are vulnerable to their own heuristics guided by decision-making processes, causing

construct-irrelevant variability in the scoring process (Wesolowski, Wind, & Engelhard, 2015). These concerns also apply to the context of music performance assessment. Specifically, music raters' decision-making processes consist of three distinct cognitive activities: (a) interpreting auditory stimuli; (b) evaluating auditory stimuli; and (c) justifying scoring decisions (Wesolowski, 2017). Errors in rater judgment (i.e., rater effects) can occur for several reasons:

> Unfortunately, the use of raters may introduce error into examinee scores for a variety of reasons—unfamiliarity with or inadequate training in the use of the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge that are relevant to making scoring decisions, or personal beliefs that conflict with the values espoused by the scoring rubric. In any case, when raters exhibit problematic rating behaviors, it may be possible to identify unique patterns in the data that correspond to specific types of rater errors (Wolfe et al., 1999, p. 4).

Rater errors can greatly influence the validity, reliability, precision, and fairness of formal performance assessments and therefore warrant serious consideration and investigation in any formalized performance assessment context, including music performance assessment.

Wesolowski et al. (2016a) described two approaches to the investigation of rater effects. The first is a rater behavior-centered approach that focuses on the ecological content of human judgment and can be classified according to four distinct areas: (a) extramusical effects related to the performer such as expressive variations (Repp, 1990, 1995), attractiveness and flair (Davidson & Coimbra, 2001; Wapnick, Darrow, Kovacs, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998), and body movement (Davidson & Correia, 2002; Davidson, 1994, 2001); (b) extramusical effects related to the assessment context such as within-ensemble communication (Wesolowski, 2013; Williamon & Davidson, 2002), acoustics (Ando, 1988), social factors (Davidson, 1997), and audience support (Berliner, 1994; Monson, 1996); (c) rater-centered effects such as memory (Radocy, 1976), first impressions (Stanley, Brooker, & Gilbert, 2002; Vasil, 1973), mood (Schubert, 1996), repertoire familiarity (Flores & Ginsburgh, 1996), and musical preference (Rentfrow & McDonald, 2009; Rentfrow, Goldberg, & Levitin, 2011; Rentfrow et al., 2012); and (d) nonmusical effects such as stereotyping (Elliott, 1995; Morrison, 1998), performance order (Bergee, 2006, 2007; Flores & Ginsburgh, 1996), evaluation time (Thompson, Williamon, & Valentine, 2007), facets of musical expression (Juslin, 2003), and teaching level and primary instrument (Hewitt & Smith, 2004). The limitation with rater behavior-centered approaches to music performance assessment protocols is that raters' observed scores are too often reported without psychometric considerations of rater behavior. Traditional indices of evaluating rater behavior in the field of music include consensus estimates of interrater reliability and consistency estimates of intrarater reliability (e.g., Bergee, 2003; Brakel, 2006; Burnsed et al., 1985; Conrad, 2003; Fiske, 1983; Hash, 2012; King & Burnsed, 2007; Norris & Borst, 2007; Silvey, 2009). The limitation with these indices when evaluating

rater behavior is that observed scores may be underestimated or overestimated if raters of varying severity/leniency rate students of the same ability (Engelhard, 1994). This effect can often present a skewed representation of what constitutes a "good," "fair," and "accurate" rater from a "bad," "unfair," and "inaccurate" rater. Wolfe et al. (2007) provide an example:

> Even when all [raters] use the scoring guidelines appropriately, traditional rater effect indices will flag some raters as exhibiting rater effects. On the other hand, if most raters are using scoring guidelines inappropriately, conventional rater effect indices will portray the best raters as outliers without indicating the higher quality of the ratings they assign (p. 2).

The second and more recent approach to investigating rater behavior in the context of music performance assessment uses empirically driven statistical indices that underscore the measurement process. Specifically, rater variability under these conditions can stem from: (a) the degree to which raters comply with the measurement instrument; (b) the way raters interpret criteria in operational scoring sessions; (c) the degree of leniency and severity exhibited; (d) raters' understanding of the measurement instrument's rating scale categories; and (e) the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks (Eckes, 2012; Wesolowski, 2017). Similar to its utility in other rater-mediated performance assessment contexts, such as writing assessments (e.g., Engelhard, 1994; Myford & Wolfe, 2003, 2004), the use of Rasch measurement models has been proven as a fruitful method for measuring latent traits mediated by raters in the context of music performance assessment (Wesolowski, Wind, & Engelhard, 2015, 2016a, 2016b). 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 rater-mediated assessments, five requirements for rater-invariant measurement underscore the Rasch measurement model: (a) rater-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular raters who happen to be used for the measuring); (b) noncrossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from raters than a less able person); (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular persons used for calibration); (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). When the data fit the requirements of the Rasch model, then rater-invariant measurement of performances is achieved (Engelhard, 2013).

Recent applications of the Rasch measurement model to music performance ratings have demonstrated a commonality of rater effects among content-experts' observed scores. In particular, rater variability in observed scores is affected by (a) rater errors, such as of severity/leniency, central tendency, halo effect, and restrictions of range (Wesolowski et al., 2016b); (b) unique interpretations of rating scale structure

(Wesolowski et al., 2016a); (c) differential rater functioning related to school levels of performances (Wesolowski et al., 2015); and (d) rater typology (Wesolowski, 2017). These investigations, however, treat rater effects as static characteristics of raters, where rater effects occur similarly across each assessed performance. With the exception of research related to high-stakes writing assessments (Hoskens & Wilson, 2001; Myford & Wolfe, 2009; Wolfe et al., 1999, 2007; Wolfe, Moulder, & Myford, 2001), the static treatment of rater effects is pervasive in research on performance assessment in general, including music performance assessment. Until now, there has been no attempt in music performance assessment contexts to systematically transform static rater measures into dynamic rater measures over time. This article presents a class of rater effects referred to as DRIFT as a mechanism to investigate rater effects as dynamic processes. The purpose of this study was (a) to evaluate the manifestation of raters' differential severity/leniency and interpretation of rating scale structure across performances and time parameters and (b) to evaluate the implications of these changes on the variability of raters' scores. The research questions that guided this study included:

1. Does the group of raters demonstrate differential severity/leniency across time points?
2. Do any individual raters demonstrate interactions between rater severity/leniency and time points?
3. Do raters systematically demonstrate differential scale category use across time points?

Using invariant measurement as a framework, this study is based on the premise that evidence of interactions between rater severity and time parameters (i.e., DRIFT) suggests that the requirements for invariant measurement are not met within an assessment system.

Although dynamic rater effects have been widely explored within the context of high-stakes writing assessments (e.g., Hoskens & Wilson, 2001; Myford & Wolfe, 2009; Wolfe et al., 1999, 2001, 2007), the application of methods for detecting rater DRIFT in other performance assessment contexts is limited. Through the application of well-established methods for exploring rater DRIFT to a new context, this study contributes to previous research on rater DRIFT in educational performance assessment in general and to research on rater effects within the context of music performance assessment in particular.

METHOD

Raters, Rating Sessions, Stimuli, and Instrument

Thirteen content-expert raters were solicited for participation in this study. The group of raters had an average of 8.25 (*SD* = 5.59) years of secondary-level instrumental teaching experience. The rating sessions occurred over the course of 5 consecutive days at the same time and in the same room for an hour and a half per session. Over the course of the 5-day rating session, a total of 75 (Day 1, *n* = 9; Day 2, *n* = 18; Day 3, *n* = 25; Day 4, *n* = 31; Day 5, *n* = 21) solo musical performances with piano accompaniment

(flute, n = 15; clarinet, n = 15; alto saxophone, n = 15; trumpet, n = 15; trombone, n = 15) were evaluated from district and state solo and ensemble performances. Thirty-seven videos represented middle school performances and thirty-eight videos represented high school performances. Acceptability of video and audio stimuli quality were previously rated and verified using the International Telecommunication Union's (2004) ITU-T rating scale. Video performances were displayed on a projector via a laptop computer with stereo sound and played repeatedly until each rater was finished responding to each item. No time limitations were placed on the rating process for each performance. Each rater used an individual laptop connected to an online response form (i.e., Google Docs) to submit ratings. The assessment design was a complete assessment network, consisting of a completely crossed two-facet design where each rater (n = 13) provided observed scores for each assessment component (i.e., rater x performance x item; Engelhard, 1997). The rating scale used was the MPR-2L-INSTSOLO (Wesolowski et. al, 2017), a 47-item Likert-type consisting of a four-point scale (e.g., strongly agree, agree, disagree, strongly disagree; see Appendix A). Prior to analysis, all data from negatively worded item stems were reverse coded to reflect similar directionality throughout.

Data Analysis Procedure: DRIFT Models

This study used a set of indicators of rater drift based on the suite of rater drift indices described in Myford and Wolfe (2009) and Wolfe et al. (2007). Specifically, three models were specified in order to explore differences in rater severity/leniency and rating scale category use across time points that suggest potential violations of the requirements for invariant measurement.

Model I. Time-static model. DRIFT refers to the changes in rater performance in relation to a parameter of time (Wolfe et al., 2001). The first model explored in this study is a version of the Multifaceted Rasch (MFR) model (Linacre, 1989) with a specific formulation to include parameters for performances, raters, items, and time (Wolfe et al., 2007). Model I provides estimates of logit-scale locations for each individual performance, rater, item, and time point. Of particular interest in the current study was the calibration of elements (i.e., days) within the time facet ($Ø_m$), which describes the average level of rater severity/leniency at each time point of interest. In order to explore the degree to which rater severity/leniency is consistent across time points, the logit-scale locations for each element within the time facet can be compared using a chi-square test.

Model II. Rater-by-time interaction model. In addition to indices of changes in overall rater severity/leniency across time points, it is also possible to explore differences in rater severity/leniency across the time periods at the individual rater level using an interaction analysis. Model II provides a means for the interpretation of rater effects as dynamic process by allowing for the evaluation of each rater's individual severity/leniency across time. Evidence of significant interactions suggests that rater calibrations are not invariant across time points.

Model III. Partial credit model for time points. The final model applied in this study was used to explore differences in raters' use of the rating scale categories across time points. Specifically, this model estimates rating scale category thresholds separately for each element of the time facet. Model III provides a means for the interpretation of rater effects as dynamic process by allowing for the comparison of the structure of the rating scale across time.

Indicators of Rater DRIFT

First, in order to explore changes in rater severity/leniency over time, the time static model (Model I) was estimated, and the logit-scale locations of the elements of the time facet were compared using a chi-square test of independence. Next, differences in rater severity/leniency across the five time points were examined using results from the interaction analysis based on Model II. Finally, differences in category use across the 5 days were examined using Model III, where the rating scale was allowed to vary across levels of the time facet.

RESULTS

Does the Group of Raters Demonstrate Differential Severity/Leniency Across Time Points?

The first research question focused on overall differences in rater severity/leniency across the five time points. Analyses related to this question were conducted using Model I (time-static model), and the corresponding results are summarized graphically in Figure 1 and statistically in Table 1.

The variable map for Model I (Figure 1) is a graphical depiction of the results from Model I that illustrates the calibrations of elements for each of the four facets on a common linear scale that represents the unidimensional latent construct. Specifically, the variable map illustrates differences in the calibrations of individual performances, raters, days, and items. The first column is the logit scale. The second column represents the spread of each of the 75 performances rated in the study. Each asterisk represents one performance, where the highest achieving performance is located at the top of the column and the lowest achieving performance is at the bottom of the column. Column 3 provides the spread of severity/leniency for each of the 13 raters. The most severe rater (Rater 5) is located at the top of the column and the least severe rater (Rater 13) is located at the bottom of the column. Column 4 represents the severity/leniency of each time point (i.e., day). Column 4 presents the calibration of the time point facet; the location of each time point (day) illustrates overall rater severity within that time point. Raters were most severe on Day 1 and least severe on Day 5. Column 5 depicts the spread of difficulty of the items. Item 30 (intonation accuracy during crescendo and decrescendo) was the most difficult item, and Item 20 (appropriate air support at various registers of the instrument) was the least difficult. Column 6 represents the

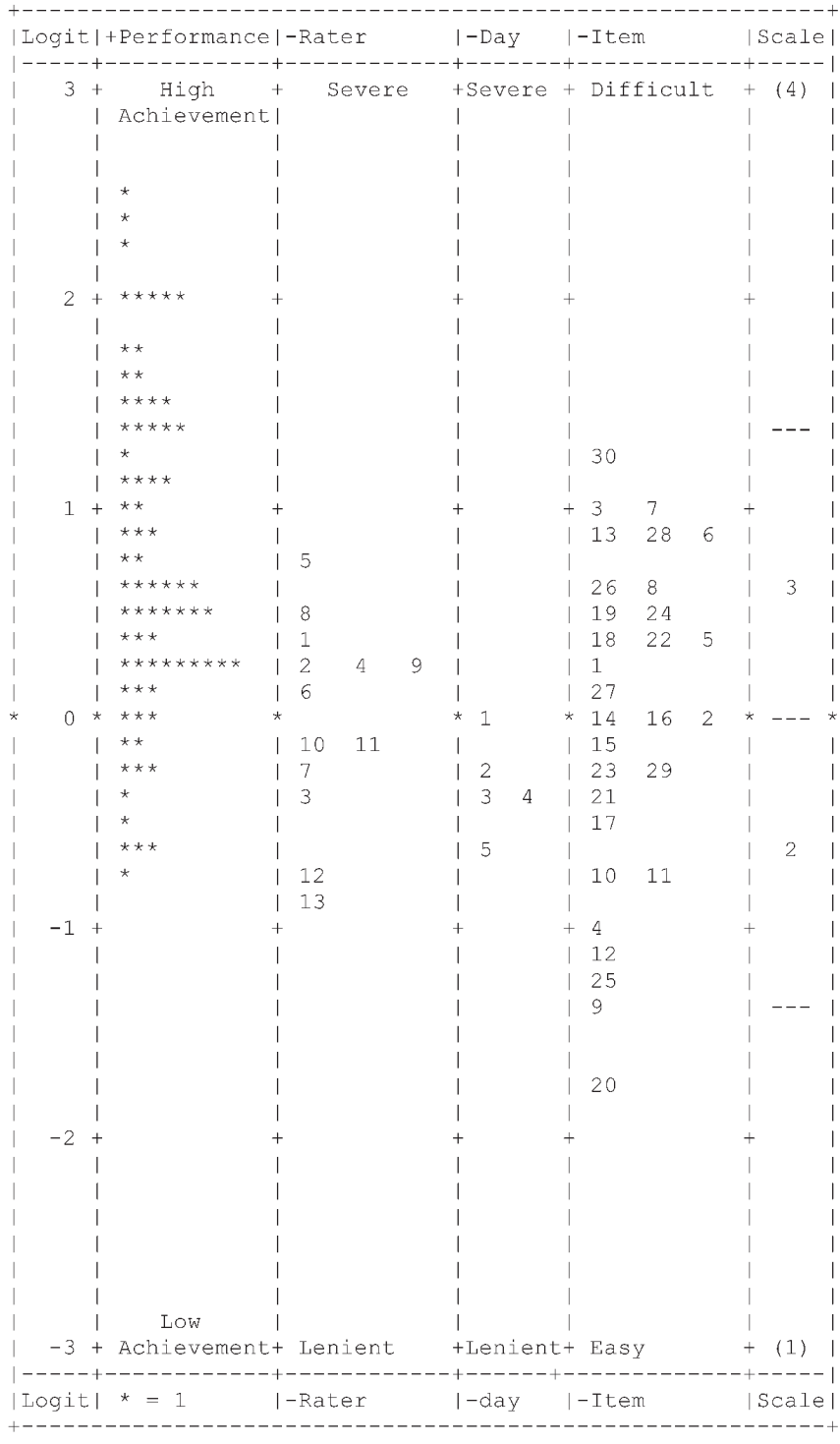

Figure 1: Variable map from Model I.

rating scale structure response format (discussed further in terms of the third research question).

Of particular interest was the calibration of the time facet. Results from the chi-square test suggested that there were significant differences among the calibrations of the 5 days on the logit scale: $\chi^2(4) = 315.1$, $p < 0.001$. This finding suggests that rater severity/leniency was significantly different between at least two of the time points examined in this study. As can be seen in the graphical display of the time point calibrations in Figure 1, along with the values in Table 1, results from Model I indicate a general trend of decreasing rater severity/leniency over the 5 days, with lower logit-scale locations corresponding to lower levels of rater severity/leniency.

In order to provide a frame of reference for interpreting the locations of elements within the time facet, the location for the first time point was fixed to zero logits. The average rater measures within each time point were as follows (see Table 1): Day 1: 0.00 logits, $n = 9$; Day 2: –0.19 logits, $n = 18$; Day 3: –0.33 logits, $n = 25$; Day 4: –0.33 logits, $n = 31$, Day 5: –0.62 logits, $n = 21$. Differences between these time points are summarized in Table 2. Substantively significant differences (logit difference > 0.30; Engelhard & Myford, 2003) were observed between Day 1 and Days 2, 4, and 5 and

Table 1
Calibration of the Time Facet (Model I)

Day	Obs. Av. Rating	Measure	*SE*	Infit	Outfit
1	2.99	0.00	0.02	1.08	1.26
2	2.93	–0.18	0.02	1.06	1.26
3	3.06	–0.32	0.02	0.95	1.05
4	3.18	–0.32	0.02	0.94	1.27
5	3.45	–0.59	0.03	1.05	1.69
Mean	3.12	–0.28	0.02	1.02	1.31
SD	0.21	0.22	0.00	0.07	0.23

Table 2
Differences in Ratings Related to Day

		Mean Differences in Achievement				
Day	Measure	1	2	3	4	5
1	0.00	—	0.19	0.33*	0.33*	0.62*
2	–0.19		—	0.14	0.14	0.43*
3	–0.33			—	0.00	0.29
4	–0.33				—	0.29
5	–0.62					
Chi-Square	*338.2					
>*df*	4					

**Note*: $p < .01$.

between Day 2 and Day 5. Each day centers close to the expected infit value of 1.00 and is within the expected infit range of .80 to 1.20, indicating that each time point demonstrated adequate fit to the model (Engelhard, 2013).

Do Any Individual Raters Demonstrate Differential Severity/Leniency Across Time Points?

The second research question focused on rater DRIFT at the individual rater level. Analyses related to this question were conducted using Model II (rater-by-time interaction model) using an interaction analysis. Results from the omnibus test revealed a significant interaction between the rater and time facets: $\chi^2(65) = 330.2$, $p < 0.001$. This finding suggests that, overall, rater severity/leniency was not invariant across the five time points.

Figure 2 presents results from the interaction analysis in terms of individual raters across the five time points. The pairwise interaction between each individual rater and the time point of interest is plotted along the *y*-axis, where different symbols are used to represent the 13 raters. Values greater than +2.00 suggest significantly higher ratings (i.e., more lenient ratings) than expected for an individual rater within a given time point, and values lower than -2.00 suggest significantly lower ratings (i.e., more severe ratings) than expected. Examination of interaction results across the 5 days indicates

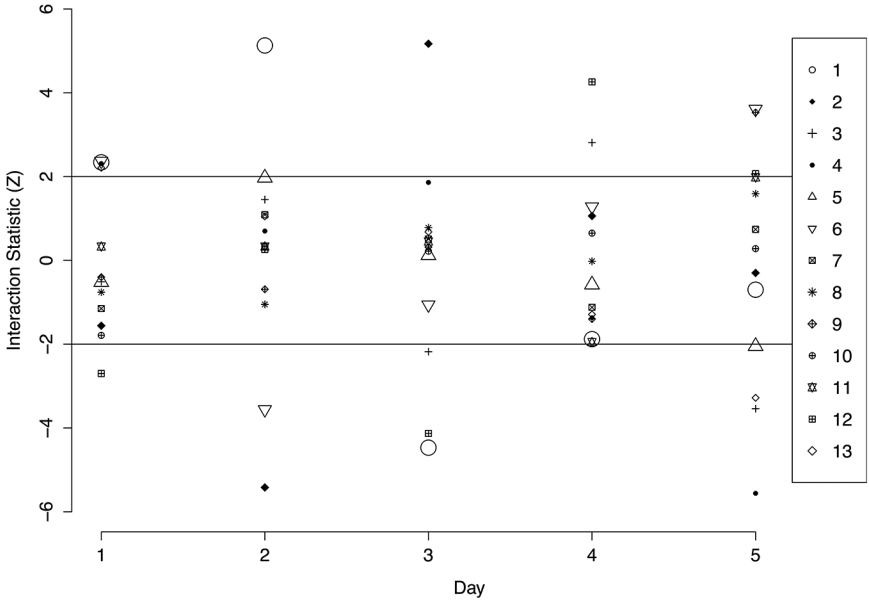

Figure 2: Rater-by-day interaction statistics.

that changes in rater severity/leniency vary in terms of direction and magnitude across the five time points.

Appendix B provides a statistical summary of the interaction results that correspond to the results illustrated in Figure 2. Of the total 65 interaction terms (13 raters/5 days), 21 (33.31%) were found to be statistically significant. Ten interactions systematically underestimated the performances and 11 interactions systematically overestimated the performances. A total of nine out of 13 raters demonstrated differential severity/leniency on at least one day. These raters included 1, 2, 3, 4, 5, 6, 9, 12, and 13. As an example, Rater 1 demonstrated systematic overestimation of performances on Day 1 (z = 2.34, observed difference = +27.17, standardized residual = 0.10) and Day 2 (z = 5.13, observed difference = +82.38, standardized residual = 0.16) but systematic underestimation of performances on Day 3 (z = 4.47 observed difference = –66.20, standardized residual = –0.15).

Do Raters Systematically Demonstrate Differential Scale Category Use Across Time Points?

The third research question for this study focused on the stability of the rating scale structure across the five time points. Analyses related to this research question were conducted using Model III. Specifically, Model III is a partial-credit (PC) formulation of the MFR model that was specified such that the structure of the rating scale was allowed to vary across time points. This model facilitated the examination of changes in rater use of the rating scale categories across the 5 days of data collection.

Figure 3 is a variable map that summarizes the results from Model III. This figure can be interpreted in a similar fashion to the variable map for Model I (Figure 1), where higher values on the logit scale indicate higher levels of achievement for the performance facet, and more severe average ratings for the rater, day, and item facets. Because of the PC formulation of the model, the variable map includes separate rating scales for each of the five time points. Specifically, separate columns are included in the variable map that illustrate the rating scale structure for the five days during which data were collected. For each day, dashed horizontal lines are used to indicate the location of rating scale category thresholds. When the PC model is used, these thresholds are the location on the logit scale at which the probability for a rating in a given category is equal to the probability for a rating in the category just below it.

The logit-scale locations for the rating scale thresholds across the 5 days are also presented in Table 3. Following the guidelines outlined by Linacre (2002), a difference in the threshold locations in the approximate range of 1.40 logits to 5.00 logits provides evidence that meaningful differences in rating scale categories exist. Examination of results from Model III suggests that meaningful differences in the structure of the rating scale across the five time points did not exist because the location of the thresholds are approximately equal across the 5 days.

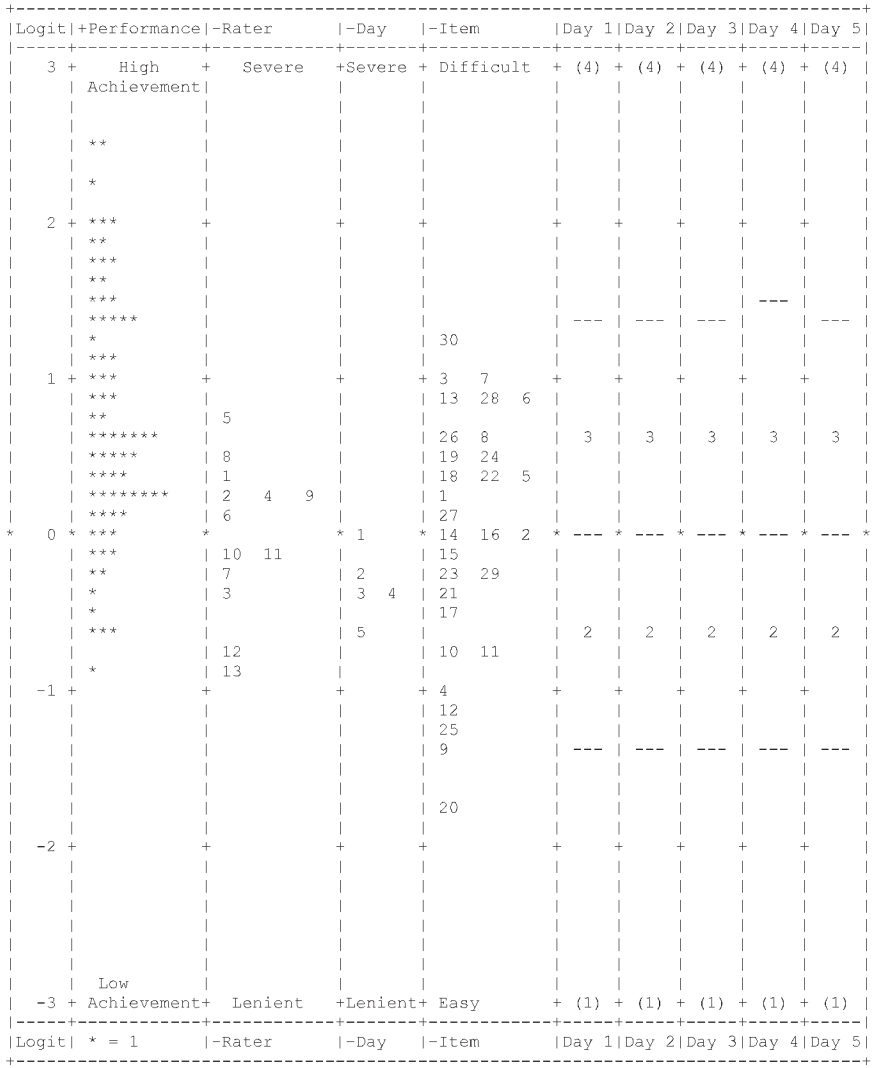

Figure 3: Variable map for Model III (partial credit model for time points).

Table 3
Rating Scale Threshold Locations Across Days (Model III)

Threshold	Day 1	Day 2	Day 3	Day 4	Day 5
τ1	–1.09	–1.14	–1.09	–1.13	–1.06
τ2	0.02	0.02	0.00	–0.02	0.00
τ3	1.06	1.11	1.08	1.14	1.05

Note: τ1 is the threshold between the "strongly disagree" and "disagree" rating scale categories; τ2 is the threshold between "disagree" and "agree" rating scale categories; τ3 is the threshold between "agree" and "strongly agree" rating scale categories.

DISCUSSION

The purpose of this study was to examine the manifestation of group and individual rater effects as well as raters' systematic changes in interpretation of a four-point rating scale structure across a 5-day rating session using the MFR model. The first research question asked if the raters, as a group, demonstrated differential severity/leniency across time points. The analysis indicated overall statistically significant differences for time points with a high reliability of separation. The raters exhibited a general trend of decreasing rater severity as the time points progressed. The second research question asked if any individual raters demonstrated differential severity/leniency across time points (i.e., interactions between rater severity/leniency and time points). The analysis indicated that nine out of the 13 individual raters exhibited differential severity/leniency during a minimum of one time point. A total of 33.31% of the pairwise interactions indicated differential severity/leniency, with 10 interactions systematically underestimating the performances and 11 interactions systematically overestimating the performances. The third research question asked if the raters systematically demonstrated a change in differential scale category use across time points. The analysis indicated that meaningful differences in the structure of the rating scale across the five time points did not exist.

The results of this study demonstrated that evidence of rater effects as dynamic processes exists in the assessment of musical performances and can negatively affect the quality of the evaluation process. As noted above, nearly all previous research related to rater DRIFT has been situated within the context of high-stakes writing assessments, where the stability of a variety of rater effects across student writing samples has been explored using methods similar to those demonstrated in the current study (e.g., Hoskens & Wilson, 2001; Myford & Wolfe, 2009; Wolfe et al., 2001, 2007). In general, these studies share a common conclusion: When examined from a dynamic, rather than static, perspective, rater effects vary across scoring periods, and rater DRIFT persists despite feedback and remediation.

The current study reflects a new application of methods for detecting rater DRIFT within the context of music performance assessment. Accordingly, the results have implications for both the "traditional" dynamic rater effects literature (high-stakes writing assessment) and the field of music performance assessment. In terms of research on writing assessment, the current results confirm those of previous studies in that differences in rater leniency/severity were observed across a scoring period—suggesting that rater DRIFT is not limited to the context of writing assessments. In terms of music performance assessment, this study illustrates methods for exploring rater effects from a dynamic perspective that provides additional insight into the quality of ratings beyond what can be observed through a static perspective. In addition to this methodological contribution, the substantive finding that the overall group of raters drifted to more lenient scoring styles as the time points progressed reflects previous research in music education. Specifically, the current findings corroborate those of Flores and Ginsburgh's (1996) related to performance order, where performances scheduled earlier in the rating

process had a lower chance of being ranked as a top performer as those who performed later in the day.

It is important to recognize that this study provides *empirical* evidence of differential severity/leniency and model-data misfit. The qualitative reasons (i.e., biases) for these systematic differences, such as fatigue, performance-to-performance carryover, response sets, and clashing standards/values, cannot be verified empirically. Millsap (2011) notes that the empirical detection of differential severity/leniency is "a device for separating the statistical phenomenon (group differences that remain after attempted matching) from what the explanation for the phenomenon might be" (p. 8). Wesolowski et al. (2015) note that the distinction between the statistical phenomenon and bias is that "bias can only be explained qualitatively through expert judgment and interpretation of systematic patterns of misfit" (p. 153). In instances when potential bias is suspected, the quantitative indices can inform us to its presence (Bond & Fox, 2015). Therefore, further investigation into the phenomena underscoring these particular results is of immediate attention and future investigation.

In a fair music performance assessment, performance ordering should not contribute to construct irrelevant variance. However, as highlighted in this study, systematic variability based upon time exists, even with content experts. As a result, it is essential that the well-established techniques for exploring rating quality from a dynamic perspective be extended to the context of music performance assessment. As a situational exemplar within the context of music education, the 2014–15 Florida Bandmasters Association's (FBA) 2014–15 District 7 Middle School Concert Music Performance Assessment was held over 3 consecutive days (Florida School Music Association, 2014). Within the consecutive 3-day rating session, raters evaluated a total of 61 middle school concert band ensembles (Day 1, n = 21; Day 2, n = 20; Day 3, n = 20). Each ensemble was slotted for 30 minutes of total stage time and performed three musical works, totaling 183 music performances (Day 1, n = 63; Day 2, n = 60; Day 3, n = 60) and 1,830 minutes (30.5 hours) of total service. According to the FBA (2015–2016) handbook, "Efforts will be made to schedule concert and jazz bands each day beginning with smaller classifications and moving in order through the larger classifications" (p. 14). With ensemble classifications based upon total school enrollment and/or level of musical repertoire, ensembles classified as High School A (total school enrollment of 2,501+, Grade 4/5-level literature) will always be scheduled at the end of the day and ensembles classified as Junior High School JC (enrollment of 1–300, Grade 1/2-level literature) will always be scheduled at the beginning of the day (FBA, 2015–2016, p. 7). As demonstrated in this study, ensembles scheduled to perform early on Day 1 (i.e., small, lower-level musical selections) would be evaluated more severely than ensembles scheduled to perform late on Day 3 (i.e., large, higher-level musical selections). In addition to these types of state- and district-based music performance assessment contexts where performance ordering is based upon classification, similar scheduling mechanisms extend to even

larger performance-based organizations such as Winter Guard International and Drum Corps International, among others.

A critical challenge in music performance assessments, such as in the FBA example, is to flag drifting raters (i.e., raters demonstrating systematic changes in severity/leniency) in real time at performance evaluations. In order to meet the need of valid, reliable, and fair assessment practice, it is first recommended that this study is followed up in a more authentic setting where raters have only one opportunity to listen to a live performance. This may provide insightful information toward rater behaviors and scale functioning. Second, it is suggested that the application of Many-Facet Rasch analyses be implemented as a reliable and systematic psychometric methodology for analyzing and reporting the quality of rater-mediated assessment data in music. Third, it is recommended that real-time analyses of rater behavior using a MFR measurement model with a time parameter be implemented and monitored by supervisory personnel trained in both music assessment and Rasch psychometric processes. Because the MFR model provides a mechanism to monitor rater behavior and detect rater errors as dynamic processes, the flagging and immediate intervention of raters demonstrating DRIFT is possible. Relying on the same judges over the course of 30.5 hours is impractical for achieving fair assessments. Strategic interventions, recalibrations, and/or replacements of raters as indicated by real-time DRIFT analysis may help improve fairness over long rating sessions.

Future research should include the investigation into the effects of rater training and rater recalibration on differential severity/leniency of dynamic processes. The benefit of DRIFT analysis is the ability to identify differential changes in rater behavior while the assessment context is occurring. If changes are identified, rater calibration protocols can be implemented in order to provide greater consistency, precision, and accuracy of ratings. These types of systems, however, have yet to be researched or implemented in the context of music performance assessment. Therefore, it is suggested that the interactions between specifically chosen music performance exemplars, types of rater training/intervention, and changes in differential severity/leniency both from a static and dynamic perspective be further studied. These investigations may include (a) how proper rater training affects changes of differential severity/leniency, (b) how the recalibration of raters through testing and retraining based upon scoring behavior of specific benchmark performances affects behavior, and (c) how the amount and type of feedback provided to raters regarding their intrarater agreement indices affects behavior. Results of such studies may provide a foundation to the establishment of proper, research-based rater calibration protocols.

Lastly, the application of DRIFT analysis may prove beneficial and as a powerful tool in the ongoing process of standardization, benchmarking, and measure construction as part of the National Association for Music Education's Model Cornerstone Assessments, the National Core Arts Music Standards development and refinement, and standardized preservice teacher evaluations such as edTPA. Because operational

raters are being used to construct measures and benchmark student performances over consecutive rating sessions, the lack of investigation toward rater behaviors over time within the context of music performance assessment, along with other performance assessment contexts, may provide construct-irrelevant variance that interferes with the validity of inferences made in the estimation of student and/or ensemble performance achievement. Using a time parameter may better inform the measurement construction process by establishing more consistent, stable, and fair scoring outcomes.

APPENDIX A

47-Item Likert-Type Scale

1.	Lack of sufficient air	SD	D	A	SA
2.	Uneven tone quality in different registers	SD	D	A	SA
3.	Appropriate dexterity when changing notes	SD	D	A	SA
4.	Intonation inaccuracy throughout the performance	SD	D	A	SA
5.	Lack of coordination between tongue and fingers/slide	SD	D	A	SA
6.	Smooth and even note changes	SD	D	A	SA
7.	Characteristic rhythmic stress of strong and weak beats	SD	D	A	SA
8.	Extraneous jaw movement	SD	D	A	SA
9.	Proper hand position	SD	D	A	SA
10.	Appropriate inflection at cadential points	SD	D	A	SA
11.	Stylistically appropriate dynamics	SD	D	A	SA
12.	Lack of meaningful contrast in dynamics	SD	D	A	SA
13.	Proper wrist position	SD	D	A	SA
14.	Proper head position	SD	D	A	SA
15.	Proper arm position	SD	D	A	SA
16.	Tone is compromised while executing expressive gestures	SD	D	A	SA
17.	Pulse and/or tempo fluctuations is stylistically characteristic	SD	D	A	SA
18.	Rhythm is not accurately subdivided	SD	D	A	SA
19.	Puffy cheeks	SD	D	A	SA
20.	Appropriate air support at various registers of the instrument	SD	D	A	SA
21.	Tempo is not appropriate	SD	D	A	SA
22.	Steady tone	SD	D	A	SA
23.	Characteristic tone	SD	D	A	SA
24.	Correct angle of instrument	SD	D	A	SA
25.	Accurately adjusts for standard instrument-related discrepancies in intonation	SD	D	A	SA
26.	Consistently set embouchure	SD	D	A	SA
27.	Elbows are pressed against the body or held out stiffly	SD	D	A	SA
28.	Characteristic embouchure	SD	D	A	SA
29.	Overblows	SD	D	A	SA
30.	Intonation accuracy during crescendo and decrescendo	SD	D	A	SA
31.	Posture exhibits tension	SD	D	A	SA
32.	Inconsistent connection of phrases to the larger context of the musical piece	SD	D	A	SA
33.	Extraneous body motion	SD	D	A	SA

34. Immediate note response in articulation	SD	D	A	SA
35. Air leaking from the corners of the mouth	SD	D	A	SA
36. Harsh and/or over tonguing in articulation	SD	D	A	SA
37. Proper breath intake before initiating tone	SD	D	A	SA
38. Consistency of articulation	SD	D	A	SA
39. Incorrect pitches	SD	D	A	SA
40. Performs with a steady pulse	SD	D	A	SA
41. Note-by-note rather than complete musical thoughts	SD	D	A	SA
42. Finger/hand tension in technical motion	SD	D	A	SA
43. Stylistically appropriate articulations	SD	D	A	SA
44. Body is slouched	SD	D	A	SA
45. Centered embouchure (left to right)	SD	D	A	SA
46. Extraneous finger/hand motion	SD	D	A	SA
47. Communication of musical phrases	SD	D	A	SA

APPENDIX B

Interaction Results for Individual Raters

Rater	Day	Infit	Outfit	Observed Total	Expected Total	Std. Mean Residual	Bias Logit	*SE*	*Z*
1	1	1.10	1.10	776.00	748.83	0.10	0.20	0.09	**2.34**
	2	1.20	1.30	1474.00	1391.62	0.16	0.33	0.06	**5.13**
	3	0.90	0.90	1203.00	1269.90	–0.15	–0.29	0.07	**–4.47**
	4	0.90	1.00	2066.00	2100.74	–0.05	–0.10	0.05	–1.88
	5	1.30	1.50	846.00	853.40	–0.03	–0.07	0.09	–0.70
2	1	1.10	1.20	755.00	775.70	–0.08	–0.12	0.07	–1.56
	2	1.00	1.20	1419.00	1523.14	–0.19	–0.28	0.05	**–5.42**
	3	1.10	1.10	1968.00	1861.00	0.17	0.26	0.05	**5.17**
	4	1.00	1.00	1975.00	1954.73	0.03	0.06	0.05	1.06
	5	1.00	1.70	611.00	613.73	–0.02	–0.03	0.11	–0.30
3	1	1.00	1.30	859.00	865.20	–0.02	–0.04	0.08	–0.51
	2	1.00	1.00	1734.00	1708.59	0.05	0.08	0.06	1.45
	3	0.90	0.90	2027.00	2067.20	–0.06	–0.12	0.05	**–2.18**
	4	0.80	0.70	2179.00	2130.99	0.08	0.17	0.06	**2.81**
	5	0.80	0.70	622.00	649.25	–0.15	–0.41	0.12	**–3.54**
4	1	1.50	1.70	800.00	770.36	0.11	0.18	0.08	2.31
	2	1.20	1.20	1077.00	1066.01	0.03	0.04	0.06	0.70
	3	1.10	1.20	1879.00	1842.04	0.06	0.10	0.05	1.86
	4	1.10	1.60	2276.00	2303.95	–0.04	–0.07	0.05	–1.37
	5	0.90	1.30	555.00	604.64	–0.28	–0.57	0.10	**–5.56**
5	1	1.00	1.10	675.00	681.17	–0.02	–0.04	0.08	–0.52
	2	1.00	1.10	1317.00	1284.95	0.06	0.12	0.06	1.97
	3	0.80	0.90	1621.00	1618.78	0.00	0.01	0.06	0.12

Rater	Day	Infit	Outfit	Observed Total	Expected Total	Std. Mean Residual	Bias Logit	*SE*	*Z*
	4	0.70	0.70	1698.00	1708.11	–0.02	–0.03	0.06	–0.58
	5	0.80	0.80	518.00	536.32	–0.10	–0.23	0.11	**–2.05**
6	1	0.90	0.80	820.00	790.21	0.11	0.19	0.08	**2.37**
	2	0.90	1.20	1404.00	1466.25	–0.12	–0.20	0.06	**–3.56**
	3	0.80	0.80	1690.00	1709.48	–0.03	–0.06	0.05	–1.06
	4	0.90	0.80	2174.00	2149.43	0.04	0.07	0.05	1.28
	5	0.90	1.10	421.00	393.69	0.23	0.54	0.15	3.61
7	1	0.90	1.10	834.00	848.46	–0.05	–0.09	0.08	–1.15
	2	1.10	1.30	1597.00	1577.79	0.04	0.06	0.06	1.09
	3	0.90	0.80	2046.00	2036.22	0.02	0.03	0.05	0.51
	4	0.90	1.60	2175.00	2195.58	–0.03	–0.06	0.05	–1.12
	5	0.90	1.40	652.00	646.09	0.03	0.10	0.13	0.74
8	1	0.90	0.90	729.00	738.86	–0.04	–0.06	0.08	–0.76
	2	1.10	1.30	1384.00	1402.65	–0.04	–0.06	0.06	–1.05
	3	1.00	1.10	1273.00	1259.67	0.03	0.05	0.06	0.78
	4	1.00	1.00	2385.00	2385.37	0.00	0.00	0.05	–0.02
	5	0.90	0.80	713.00	697.09	0.08	0.17	0.10	1.59
9	1	1.50	2.10	763.00	768.30	–0.02	–0.03	0.08	–0.40
	2	1.20	1.60	1049.00	1060.13	–0.03	–0.04	0.06	–0.69
	3	1.20	1.40	2220.00	2212.80	0.01	0.01	0.05	0.32
	4	1.00	1.10	1477.00	1500.13	–0.05	–0.08	0.06	–1.40
	5	1.20	1.40	641.00	608.98	0.18	0.44	0.12	**3.53**
10	1	0.80	0.80	808.00	829.13	–0.08	–0.15	0.08	–1.79
	2	0.70	0.80	1163.00	1158.01	0.01	0.02	0.07	0.34
	3	0.80	0.80	2370.00	2365.77	0.01	0.01	0.05	0.22
	4	0.90	0.80	1586.00	1576.61	0.02	0.05	0.07	0.65
	5	1.00	1.10	628.00	625.78	0.01	0.04	0.13	0.28
11	1	0.90	1.00	830.00	825.54	0.02	0.02	0.07	0.33
	2	1.30	2.10	1539.00	1532.63	0.01	0.02	0.05	0.33
	3	0.80	1.70	1999.00	1989.64	0.01	0.02	0.05	0.45
	4	1.10	1.90	2034.00	2071.28	–0.06	–0.10	0.05	–1.95
	5	1.00	4.60	660.00	643.18	0.09	0.25	0.13	1.98
12	1	1.20	1.50	886.00	918.19	–0.12	–0.22	0.08	**–2.70**
	2	1.00	0.90	1293.00	1289.10	0.01	0.02	0.07	0.26
	3	1.10	1.20	1502.00	1563.02	–0.14	–0.26	0.06	**–4.13**
	4	1.20	2.60	1892.00	1831.39	0.12	0.34	0.08	**4.26**
	5	1.10	2.60	2107.00	2077.67	0.05	0.16	0.08	**2.07**
13	1	1.10	1.50	957.00	932.06	0.09	0.21	0.09	**2.21**
	2	0.80	1.10	1757.00	1740.59	0.03	0.07	0.07	1.05
	3	0.70	0.70	1591.00	1581.60	0.02	0.05	0.07	0.68
	4	0.70	1.70	2471.00	2491.29	–0.03	–0.08	0.06	–1.29
	5	0.90	1.70	1078.00	1108.15	–0.10	–0.32	0.10	**–3.28**

Note: Shaded rows indicate $|Z| > 2.00$.

REFERENCES

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

Ando, Y. (1988). *Architectural acoustics: Blending sound sources, sound fields, and listeners*. New York, NY: Springer.

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

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

Barnes, G. V., & McCashin, R. (2005). Practices and procedures in state adjudicated orchestra festivals. *Update: Applications of Research in Music Education, 23*(2), 34–41.

Bergee, M. J. (1989). An investigation into the efficacy of using an objectively constructed rating scale for the evaluation of university-level single-reed juries. *Missouri Journal of Research in Music Education, 26*, 74–91.

Bergee, M. J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education, 45*(4), 601–612.

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

Bergee, M. J. (2006). Validation of a model of extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 54*(3), 244–256.

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55*(4), 344–358.

Berliner, P. J. (1994). *Thinking in jazz: The infinite art of improvisation*. Chicago, IL: University of Chicago Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: 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.

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

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

Davidson, J. W. (1994). Which areas of a pianist's body convey information about expressive intention to an audience? *Journal of Human Movement Studies, 26*, 279–301.

Davidson, J. W. (1997). The social in musical performance. In D. J. Hargraves & A. C. North (Eds.), *The social psychology of music* (pp. 209–228). Oxford, England: Oxford University Press.

Davidson, J. W. (2001). The role of the body in the production and perception of solo vocal performance: A case study of Annie Lennox. *Musicae Scientiae, 5*(2), 235–256.

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.

Davidson, J. W., & Correia, J. S. (2002). Body movement. In R. Parncutt & G. E. McPherson (Eds.), *The science and psychology of music performance* (pp. 237–249). Oxford, England: Oxford University Press.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292.

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education, 127*, 50–56.

Ellis, M. C. (1997). An analysis of taped comments from a high school jazz band festival. *Contributions to Music Education, 34*, 35–49.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 33*(2), 115–116.

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

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

Engelhard, G., & Myford, C. M. (2003). *Consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model* (Report No. 2003–1). New York, NY: College Board.

Fautley, M. (2010). *Assessment in music education*. New York, NY: Oxford University Press.

Fiske, H. E. (1983). *The effect of a training procedure in music performance evaluation on judge reliability*. Toronto, ON: Educational Research Council.

Flores, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking. *The Statistician, 45*(1), 97–104. doi:10.2307/2348415

Florida Bandmasters Association. (2015–2016). Florida Bandmasters Association *2015–2016 handbook*. Retrieved from http://fba.flmusiced.org/media/1279/handbook15-16.pdf

Franklin, J. O. (1979). *Attitudes of school administrators, band directors, and band students towards selected activities of the public school band program (Doctoral dissertation)*. Available from Proquest Dissertations and Theses database. (UMI 8007218)

Florida School Music Association. (2014). *FBA MPA programs and results*. Retrieved from http://flmusiced.org/mpaonline/publicreports/mpamenu.aspx?ComponentID=1

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.

Hewitt, M. P., & Smith, B. P. (2004). The influence of teaching-career level and primary performance instrument on the assessment of music performance. *Journal of Research in Music Education, 52*(4), 314–327.

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

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*, 121–146.

Howard, K. K. (1994). *A survey of Iowa high school band students' self-perceptions and attitudes toward types of music contests (Doctoral dissertation)*. University of Iowa, Iowa City.

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

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

International Telecommunication Union. (2004). *Objective perceptual assessment of video quality: Full reference television*. *ITU-T Telecommunication Standardization Bureau*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Objective+perceptual+assessment+of+video+quality+:+Full+reference+television#6

Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music, 31*(3), 273–302. doi:10.1177/03057356030313003

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, 45*(1), 27–33.

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

Kruth, E. (1970). Adjudication and the music festival. *Instrumentalist, 24*(6), 48.

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.

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). *Many facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

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, England: 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.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Monson, I. (1996). *Saying something: Jazz improvisation and interaction*. Chicago, IL: University of Chicago Press.

Morrison, S. J. (1998). A comparison of reference responses of White and African-American students to musical versus musical/visual stimuli. *Journal of Research in Music Education, 46*, 208–222.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using the many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

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

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389.

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*(3), 237–251.

Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education, 24*, 119–128.

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology, 100*(6), 1139–1157.

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The song remains the same: A replication and extension of the MUSIC Model. *Music Perception, 30*(2), 161–185.

Rentfrow, P. J., & McDonald, J. A. (2009). Preference, personality, and emotion. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, application* (pp. 669–695). Oxford, England: Oxford University Press.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*(2), 622–641.

Repp, B. H. (1995). Expressive timing in Schumann's "Träumerei:" An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America, 98*(5), 2413–2427.

Schubert, E. (1996). Enjoyment of negative emotions in music: An associative network explanation. *Psychology of Music, 24*, 18–28.

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.

Sweeney, C. R. (1998). *A description of student and band director attitudes toward concert band competition (Doctoral dissertation)*. University of Miami, Coral Gables, FL.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception, 25*(1), 13–29.

Vasil, T. (1973). *The effects of systematically varying selected factors on music performing adjudication (Doctoral dissertation)*. University of Connecticut, Storrs.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education, 45*(3), 470–479.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on evaluation of violin performance evaluation. *Journal of Research in Music Education, 46*(4), 510–521.

Wesolowski, B. C. (2013). Cognition and the assessment of interaction episodes in jazz improvisation. *Psychomusicology: Music, Mind, and Brain, 23*(4), 236–242.

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

Wesolowski, B. C., Amend, R., Barnstead, T., Edwards, A., Everhart, M., Goins, Q., . . . Williams, J. (2017). The development of a secondary-level solo wind instrument performance rubric using the Multifaceted Rasch partial credit measurement model. *Journal of Research in Music Education, 65*(1), 95–119.

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

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Examining rater precision in music performance assessment: An analysis of rating scale structure using the Multifaceted Rasch partial credit model. *Music Perception, 33*(5), 662–678.

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

Williamon, A., & Davidson, J. W. (2002). Exploring co-performer communication. *Musicae Scientiae, 5*(1), 53–72.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83–106.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (1999). *Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*(3), 256–280.

Wolfe, E. W., Myford, C. M., Engelhard Jr., G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays* (Report No. 2007–2). New York, NY: College Board.