Bulletin of the Council for Research in Music Education
Winter 2018 No. 215

Evaluating Differential Rater Accuracy Over Time in Solo Music Performance Assessment

Stefanie A. Wind
The University of Alabama
Tuscaloosa, AL

Brian C. Wesolowski
The University of Georgia
Athens, GA

ABSTRACT

In formal music performance assessments where raters evaluate performances by different students or ensembles over several days, the quality of raters' judgments can vary over the duration of the assessment. Because time is construct-irrelevant, the influence of time on rater judgments poses a threat to the fairness of the assessment. In this study, we used Rasch measurement theory to explore changes in rater accuracy over time in the context of a secondary-level solo instrumental music performance assessment, where rater accuracy is defined as the match between observed ratings and criterion ratings. Specifically, we considered changes in rater accuracy related to the overall group of raters, individual raters, and rubric domains. Results suggested that the raters became less accurate over time, and there were differences in rater accuracy within domains over time. Together, our findings highlight the dynamic nature of rater accuracy over time. Our findings also highlight the importance of empirically examining changes in rater accuracy over time using the approach illustrated in this study. Implications for investigating rater accuracy over time in both research-based and applied assessment contexts are discussed.

In any formal music performance assessment where judges (i.e., raters) are used to operationally score performances, an opportunity is allowed to introduce construct-irrelevant variability into the assessment process. Empirical evidence in prior research indicates that in the context of music performance assessments, rater variability can stem from raters' leniency/severity (Wesolowski, Wind, & Engelhard, 2016a), the precision in raters' use of a measurement instrument's rating scale structure (Wesolowski, Wind, & Engelhard, 2016b), raters' systematic differential severity based upon the performers' subgroup affiliation (Wesolowski, Wind, & Engelhard, 2015), and raters' differential severity of item use due to personal idiosyncrasies defined by rater type (Wesolowski, 2017). In each of these studies, rater variability was interpreted as a *static* characteristic of the rater, where the effect of the rater was

treated equally across each musical performance. However, in formal music performance assessments, it is common for raters to evaluate performances over the course of several days, and these days can last between 8–16 hours (Barnes & McCashin, 2005). In these instances, rater variability can be considered a changing, or *dynamic*, characteristic of the rater, where rater effects can enter the music assessment context systematically as differential functioning over time (Wesolowski, Wind, & Engelhard, in press). Raters who demonstrate systematic differences over time in leniency (i.e., the systematic raising of ratings during the course of a scoring session) or severity (i.e., the systematic lowering of ratings during the course of a scoring session) contribute a different type of construct-irrelevant variance to the assessment context that is not often considered in the research literature or in applied assessment contexts, thereby obscuring the validity, reliability, and, most importantly, fairness of the assessment context (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Therefore, one major challenge of managing the fairness of formal music performance assessments is the ability to identify raters who demonstrate changes in their rating patterns over the course of time.

Fairness is a validity issue concerned with the degree to which measurement procedures result in accurate estimates of student achievement in terms of a construct (Wesolowski & Wind, in press). The recent revision of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) indicates that "regardless of the purpose of testing, the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure" (p. 51). The *Standards* further indicate that "a fair test does not advantage or disadvantage some individuals because of the characteristics irrelevant to the intended construct" (AERA, APA, & NCME, 2014, p. 50). For music performance assessments, one such characteristic is time. Therefore, the consideration of time effects in the context of a rater-mediated music performance assessment is critical to the fairness of any music performance assessment. Dynamic rater effects have been examined in detail within the context of writing assessment using the many facet Rasch (MFR) measurement model containing a time facet (e.g., Hoskens & Wilson, 2001; Myford & Wolfe, 2009; Wolfe, 2004; Wolfe, Moulder, & Myford, 2001; Wolfe, Myford, Engelhard, & Manalo, 2007). Time facet variations of the MFR model allow for the investigation of rater errors over time, or more specifically described, differential rater functioning over time (also commonly referred to as *rater DRIFT*). Rater DRIFT is not discoverable using a static version of the MFR measurement model. According to Wolfe et al. (2001), differential rater functioning over time can manifest itself in several ways: (a) primacy/recency (i.e., *primacy* raters assign higher ratings to performances evaluated earlier in the assessment context or *recency* raters assign higher ratings to performances evaluated later in the assessment context), (b) practice/fatigue (i.e., *practice* raters demonstrate higher levels of agreement over the course of the assessment context or *fatigue* raters demonstrate lower levels of agree-

Table 1
Rating Quality Indices Based Upon the MFR Model for Rater Accuracy

Category	Indicators and displays based on the MFR model	Substantive interpretation (questions)	Statistics and displays
A. Rater accuracy calibrations	1. Rater leniency/severity accuracy 2. Rater accuracy precision 3. Rater accuracy separation	1. What is the accuracy location of each rater? 2. How precisely has each rater been calibrated in terms of accuracy? 3.1. How spread out are the individual raters in terms of accuracy? 3.2. Can the raters considered to be exchangeable in terms of accuracy?	1a. Variable map 1b. Calibration and location of elements within facet 2. Standard errors for raters 3.1. Reliability of separation statistic for raters 3.2. Chi-square statistic for raters
B. Model-data fit	1. Model-data fit for rater accuracy	1. How consistently does each rater demonstrate accuracy across domains, rating scale categories, and/or musical performances?	1. Mean square error fit statistics (infit and outfit mean square error)
C. Interactions	1. Rater accuracy interactions	1. Is rater accuracy invariant across the facets included in the interaction term? 2. Is rater accuracy invariant within the pairs of individual elements included in the interaction term?	1. Omnibus test for the interaction between two facets 2. Pairwise interaction terms

Note: Adapted from Wind and Engelhard (2013) and Wesolowski and Wind (2017).

ment over the course of the assessment context), and (c) differential centrality/differential extremism (i.e., *differential centrality* raters tend to use more of the rating categories in the center of the rating scale over the course of the assessment context or *differential extremism* raters tend to use more of the rating categories at the extremes of the rating scale over the course of the assessment context). In this study, we explore *differential accuracy* over time in order to identify how DRIFT may occur in the context of a secondary-level solo instrumental music performance assessment, where the alignment between observed rat-

ings from operational raters and criterion scores established as accurate (i.e., "true scores") changes over the course of the assessment (discussed later in this article). Wolfe et al.'s examples can serve as possible explanations for DRIFT if empirical evidence of DRIFT emerges in the analysis. Therefore, once empirical evidence of DRIFT is identified, each of Wolfe's possible explanations can be considered, post hoc, as a potential cause based upon a qualitative investigation of the observed scores. Regardless of the specific type of DRIFT, all of these manifestations of differences in rater judgment involve the influence of time as a form of construct-irrelevant variance in rater judgments. As a result, all forms of DRIFT pose a threat to the procedural fairness and, ultimately, the validity of a rater-mediated performance assessment (Kane, 2010).

Rater DRIFT has most often been described in the context of rater-mediated writing assessments in which raters score essays over multiple days. However, various forms of rater DRIFT have also been described in performance assessments in economics (Hoskens & Wilson, 2001) and in performance assessment generally using simulated data (e.g., Wolfe et al., 2001). In the context of music assessment, a recent study by Wesolowski et al. (in press) provided empirical evidence that in the context of a formal, 5-day music performance assessment, raters demonstrated systematic differences in severity based upon a time parameter demarcated by day. More specifically, it was found that, overall, raters demonstrated a general trend of decreasing severity (i.e., the systematic raising of ratings) over a 5-day rating session. According to Wolfe et al.'s (2001) description, overall the raters demonstrated differential leniency in the form of recency. Hypothetically, then, the same performance that was rated earlier in the scoring period would have received lower scores than if it was rated later in the scoring period. As a result, according to the definition provided in the *Standards* (AERA, APA, & NCME, 2014), the assessment context evaluated in Wesolowski et al.'s (in press) study was unfair due to the performances being scored differently based specifically upon changes in scoring over the course of the 5-day rating session.

Rater DRIFT is a serious threat to the validity of any rater-mediated music performance assessment. As an independent analysis like the study by Wesolowski et al. (in press) describes, rater DRIFT analysis provides validity evidence for the assessment context by identifying differential severity across time. However, it does not provide diagnostic information for the interpretation and lessening of the differential severity. As a result, methodological steps are needed to move beyond that of simply identifying it. One methodology that allows for the correction of rater DRIFT is the concurrent implementation of *rater accuracy* indices to DRIFT analysis, where rater accuracy is defined as the degree to which operational raters' ratings match criterion scores (i.e., true scores or expert scores) on the same performances. Wesolowski and Wind (2017) demonstrated that rater accuracy indices, as an independent analysis in the context of static rater analysis, are a beneficial means to provide concrete evidence of the specific manner in which the rater applies the measurement instrument in the context of a solo music performance assessment. Specifically, the study used criterion-referenced accuracy indicators based on Rasch measurement theory, which provided a meaningful frame of reference for considering

rating quality based on the operational scoring of an expert rater. The Rasch approach to exploring rater accuracy involves specifying a model in which rater accuracy, as defined by the match between operational and expert raters, is the latent variable (i.e., construct) on which raters, performances, and other researcher-specified variables (i.e., facets) are calibrated. The model results in estimates on an interval-level scale (the log-odds, or "logit" scale) that reflect the difficulty associated with providing accurate ratings for individual raters, performances, and other facets. Specifically, this scale reflects the odds of an accurate rating, rather than an inaccurate rating, given a particular combination of raters, performances, and other facets, such as items.

Using the Rasch measurement approach to evaluating rater accuracy, three major categories of rater accuracy indicators can be examined. Table 1 includes a summary of three categories of rater accuracy indices based on Rasch measurement theory: (a) rater accuracy calibrations, (b) model-data fit, and (c) interactions. Each of these categories of rater accuracy indices reflects a different aspect of rater accuracy based on the match between operational and expert ratings. The first set of indices is based on rater accuracy calibrations. These indicators include numeric estimates of each rater's location on the latent variable that reflects rater accuracy, where higher locations indicate that a rater is more accurate and lower locations indicate that a rater is less accurate. Additional indicators within the first category include estimates of the precision of the location estimates for each rater in the form of standard errors, along with indicators of the degree to which raters' accuracy locations are significantly different. The second set of indices is based on model-data fit statistics. These statistics are numeric indicators that describe the degree to which rater accuracy calibrations can be interpreted as their locations on the latent variable that represents rater accuracy. Finally, interactions can be examined within the framework of rater accuracy. When the Rasch approach to evaluating rater accuracy is employed, interaction analyses provide insight into the degree to which there are systematic differences in rater accuracy across facets in the assessment procedure, such as domains and days in the assessment procedure. Significant interactions suggest that rater accuracy is not invariant across individual domains, days, or levels of other facets in the assessment system.

Based on these Rasch indicators of rater accuracy, Wesolowski and Wind's (2017) study provided: (a) rater accuracy indices for the overall accuracy of raters throughout the assessment, (b) the accuracy of raters across domains of the scoring rubric, and (c) the accuracy of raters across items of the scoring rubric. The benefit of obtaining accuracy indices in music performance assessment is that it allows the opportunity to employ rater-training procedures that can ultimately improve the psychometric soundness (i.e., validity, reliability, and fairness) of music performance assessments.

We propose in this current study that the consideration of criterion-referenced rater accuracy indicators based on Rasch measurement theory, conceptualized dynamically by specifically including them within a time facet, can provide meaningful reporting data as to how a rater applies the measurement instrument throughout the duration of a music

performance assessment. Using the definition of rater accuracy as the degree to which operational raters' ratings match criterion scores on the same performances, rater accuracy can be viewed as a criterion-referenced indicator of rating quality that provides evidence for the validity and fairness of a rater-mediated assessment. Because time is a construct-irrelevant characteristic in rater-mediated performance assessments, systematic changes in rater accuracy related to time pose a threat to the fairness of the assessment.

Accordingly, the most meaningful indices from Table 1 for the purpose of exploring rater accuracy DRIFT are interaction terms between time and other facets in the model. These interaction terms provide evidence regarding the degree to which rater accuracy is invariant over the course of a multiday assessment. Additional analyses related to observed ratings, including analyses specifically related to rater errors and systematic biases and reliability, as well as qualitative analyses are needed in order to more fully understand and explain changes in rater accuracy over time.

PURPOSE

The purpose of this study is to explore differences in rater accuracy over time within the context of a solo music performance assessment, where rater accuracy is defined as the match between operational raters' ratings and criterion scores. Specifically, we considered differences in rater accuracy over 5 days of a solo instrumental music performance assessment as they relate to the overall group of raters, individual raters, and rater accuracy within the domains in the scoring rubric. Three research questions guided the analyses:

1. Are there overall differences in rater accuracy across the duration of the music solo performance assessment?
2. Do any individual raters demonstrate interactions between accuracy and days of the assessment?
3. Is rater accuracy within domains invariant across days of the assessment?

This study contributes to research on music performance assessments in two main ways. First, it highlights the dynamic nature of rater accuracy in multiday rater-mediated music performance assessments. Accordingly, this study emphasizes the importance of evaluating differences in rater accuracy over time as a key component of validity evidence for these assessments. Second, this study has methodological implications. Specifically, the analysis in this study demonstrates a procedure for evaluating rater accuracy from a dynamic perspective that can be applied to multiday music performance assessments beyond the specific assessment examined in the current study.

METHODS

Instrument

The measurement instrument used in this study was the Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO; Wesolowski et al., 2017; see Figure 1 for items and domains). The rubric consisted of a total of 28 items. Each item

was scored using a rating scale that included between two and four rating scale categories (see Appendix A here: http://bcrme.press.illinois.edu/media/215/). The items were positioned within eight domains: (a) Technique (n = 2), (b) Tone (n = 2), (c) Articulation (n = 1), (d) Intonation (n = 1), (e) Visual (n = 9), (f) Air Support (n = 3), (g) Melody (n = 4), and (h) Expressive Devices (n = 6).

Domain	Items
A. Technique	1. Finger/slide dexterity
	2. Coordination between tongue and fingers/slide
B. Tone	3. Tone quality in varying registers
	4. Tone while executing expressive gestures
C. Articulation	5. Consistency of articulation
D. Intonation	6. Intonation accuracy
E. Visual	7. Body posture
	8. Instrument angle
	9. Head position
	10. Arm position
	11. Wrist position
	12. Hand position
	13. Embouchure/flexibility
	14. Cheeks
	15. Jaw movement
F. Air support	16. Breath intake
	17. Sufficiency of air
	18. Air support in various registers of the instrument
G. Melody	19. Note accuracy
	20. Communication of musical phrases
	21. Connection of Phrases
	22. Inflection at cadence points
H. Expressive Devices	23. Stylistically related dynamics
	24. Contrast in dynamics
	25. Subdivision of the rhythm
	26. Appropriateness of tempo
	27. Steadiness of pulse
	28. Expressive pulse and tempo fluctuations

Figure 1. Assessment items from the Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO; Wesolowski et al., 2017; Wesolowski et al., in press).

Performance stimuli. A total of 89 secondary-level (e.g., grades 6–12) video musical performances were evaluated for this study: flute (n = 18), clarinet (n = 17), saxophone (n = 11), oboe (n = 6), bassoon (n = 4), trumpet (n = 7), trombone (n = 12), French horn (n = 7), euphonium (n = 3), and tuba (n = 3). The sample was gathered on a volunteer basis from two suburban schools (high school, n = 1; middle school, n =1). Participants had an average of 3.21 (SD = 1.91) years of performance experience, and the performance quality represented an average quality of performance achievement for secondary-level (grades 6–12) solo music performances. The participants were allowed to select their own piece of music with no requirements from the researchers or participating teacher. Acceptability of video and audio stimuli quality were previously rated and verified using the International Telecommunication Union's ITU-T Rating Scale (Union, 2004). Each of the performances was displayed onto a projector from a single laptop along with stereo sound. All performances were recorded with the student in a seated position, and they read their selected piece of music on a black Manhasset music stand. No student performances were affiliated with any of the rater participants in this study. All students and legal guardians completed documentation of informed consent.

Raters. A total of eight operational raters and one expert rater evaluated each of the 89 musical performances; all of the raters evaluated all of the performances, such that the rating design was fully crossed. The operational raters were defined as subject matter experts in this study, as their teaching background and experience were favorably representative of the performance stimuli (e.g., secondary-level instrumental performances) being evaluated. The eight operational raters represented varied demographics: teaching locales (urban, n = 4; suburban, n = 3; rural, n = 1), current teaching level (middle school, n = 4; high school, n = 3; collegiate, n = 1); years of teaching experience (M = 7.78, SD = 4.21), minimum degree level (bachelor's, n = 4; master's, n = 4), and primary instrument (woodwind, n = 4; brass, n = 4). The expert rater had expertise in the research area of measurement and evaluation and also played a fundamental role in the development of the MPR-2L-INSTSOLO. Because this rater was directly involved in the development of this instrument, the rater was considered an expert in the construct defined by the rubric. Empirical evidence to support the interpretation of this rater's judgments as meaningful criteria was also gathered in a previous analysis of the observed ratings (Wesolowski et al., 2017). Specifically, previous analyses revealed that the expert's ratings demonstrated desirable psychometric properties, including moderate severity, effective use of the rating scale categories, and adequate model-data fit to the Rasch model. Accordingly, this rater's judgments can be viewed as meaningful criterion against which to compare operational rater judgments that is congruent with the selection of expert raters in previous accuracy studies (e.g., Engelhard, 1996; Wesolowski & Wind, 2017; Wesolowski, Wind, & Engelhard, in press; Wind & Engelhard, 2013; Wolfe, Song, & Jiao, 2016). A complete assessment network was used in this study, where all raters evaluated all musical performances (Engelhard, 1997) over the 5-day period in the same room, at the same time, and

with no consultation or discussion of ratings with each other. The performance ordering was selected randomly based upon a nominal identifier.

Data analysis. In order to explore the research questions for this study, it was necessary to calculate accuracy scores for each of the operational raters' ratings of the student performances. Following Engelhard (1996), rater accuracy scores were calculated by comparing each operational rater's rating to that of the expert rater on the same performance and item. An accuracy score of "1" was assigned when there was an exact match between the operational and expert rater, and an accuracy score of "0" was assigned otherwise. Using these dichotomous accuracy scores, three formulations of the many-facet Rasch rater accuracy (MFR-RA) model were applied that reflect the three guiding research questions for this study. All of the Rasch analyses were conducted using the Facets computer program (Linacre, 2015).

Model I. The first model is a dichotomous MFR-RA model that provides an overall summary of rater accuracy across raters, performances, items, domains, and days in the solo music performance assessment. The model is stated mathematically as:

$$\left[\frac{P_{ijmkn(x=1)}}{P_{ijmkn(x=0)}}\right] = \lambda_i - \beta_j - \delta_m - \eta_k - \gamma_n \qquad (1)$$

where

$P_{ijmk(x=1)}/P_{ijmk(x=0)}$ = the probability that rater i provides an accurate rating ($x = 1$), rather than an inaccurate performance ($x = 0$) on student performance j on item m within domain k on day n.

λ_i = the ability of rater i to provide accurate ratings;

β_j = the difficulty associated with providing an accurate rating to student performance j;

δ_m = the difficulty associated with providing an accurate rating on item m;

η_k = the difficulty associated with providing an accurate rating on domain k; and

γ_n = the difficulty associated with providing an accurate rating on day n.

The outcome variable in Equation 1 is expressed as the log-odds of an accurate rating rather than an inaccurate rating, given a particular combination of an individual rater, a specific student performance, a particular item, a specific domain, and a particular day in the multiday music performance assessment. This log-odds (i.e., logit) scale is also used to describe the overall accuracy for each rater (λ), student performance (β), item, domain (η), and day (γ). As a result, the common linear metric that represents rater accuracy, individual raters, performances, items, domains, and days can be compared.

When Model I is applied to dichotomous accuracy scores, estimates are obtained for each rater, student performance, item, domain, and day. Differences among individual rater locations reflect differences in the overall accuracy level of each rater who scored the solo music performance assessment. Locations within the remaining facets reflect the overall level of rater accuracy across individual student performances, items, domains, and days in the solo music performance assessment.

Model II. The second model is used to explore interactions between individual raters and days of the assessment procedure. Results from this model provide insight into the degree to which individual raters' accuracy changed across the 5 days of the solo music performance assessment. Model II has the same basic structure as Model I but includes an interaction term between the rater and day facets:

$$\ln\left[\frac{P_{ijmkn(x=1)}}{P_{ijmkn(x=0)}}\right] = \lambda_i - \beta_j - \delta_m - \eta_k - \gamma_n - (\lambda_i\gamma_n) \qquad (2)$$

In Model II, the term $\lambda_i\gamma_n$ is used to investigate the null hypothesis that individual raters' accuracy levels are invariant across each of the days of the assessment procedure. Results from the interaction analysis are evaluated using an overall omnibus test for the interaction term as well as pairwise tests for interactions between each rater and each day.

Model III. Finally, Model III is used to explore interactions between domains on the assessment rubric and days of the assessment procedure. Results from this model provide insight into the degree to which overall rater accuracy within domains varied across days of the music performance assessment. Accordingly, this model includes the same basic structure as Model I, with the addition of an interaction term between the domain and day facets:

$$\ln\left[\frac{P_{ijmkn(x=1)}}{P_{ijmkn(x=0)}}\right] = \lambda_i - \beta_j - \delta_m - \eta_k - \gamma_n - (\eta_k\gamma_n) \qquad (3)$$

Similar to Model II, the term $\eta_k\gamma_n$ is used to investigate the null hypothesis that rater accuracy within domains is invariant across each of the days of the assessment procedure. Results from the interaction analysis are evaluated using an overall omnibus test for the interaction term as well as pairwise tests for interactions between each domain and each day.

RESULTS

Model I

Summary statistics. Table 2 includes summary statistics from Model I (Equation 1). As noted above, this model provides an overall summary of rater accuracy in scoring the solo music performance assessment in terms of student performances, items, domains, and days of the assessment. For each facet, the average logit-scale location is reported along with indicators of model-data fit and separation. In the context of the current study, logit-scale locations for each facet represent their accuracy measure, calculated using Equation 1. For raters, the logit-scale locations reflect the overall accuracy of each rater across the 89 performances and 28 items. Higher accuracy measures for raters suggest that a rater was accurate more often. For items and domains, the accuracy measure

indicates the difficulty associated with assigning an accurate rating for a particular item or domain. Higher measures suggest that the item or domain was difficult to rate accurately. For days, accuracy measures reflect raters' overall accuracy within the day of the music performance assessment. Higher measures suggest that raters were overall more accurate on the particular day. In order to provide a frame of reference for interpreting the rater accuracy locations on the logit scale, all of the facets except the rater facet were centered (mean set to zero), and the rater facet was allowed to vary. The results from Model I indicate that the average rater location on the logit scale is higher (M = 0.26, SE = 0.10) than average student performance, item, domain, and day locations. The finding that the raters were located higher on the logit scale compared to the other facets suggests that the raters who scored the solo music performance assessment were generally accurate across the performances, items, domains, and days of the assessment.

In terms of model-data fit, values of the infit *MSE* and outfit *MSE* statistics for all of the facets suggest acceptable fit between the MFR-RA model and the rater accuracy scores examined in this study. Specifically, the values of the *MSE* fit statistics are close to 1.00, which is the generally accepted expected value of these statistics when data fit the model (Engelhard, 2013). This result suggests that patterns of raters' accuracy scores across the performances, items, domains, and days did not deviate substantially from what would be expected if the data fit the model perfectly. Accordingly, the logit-scale locations based on Model I can be interpreted as indicators of rater, student performance, item, domain, and day locations on the latent variable that represents rater accuracy. Finally, the separation statistics for each of the facets in Model I suggest that there are significant differences (p < 0.01) in rater accuracy based on the chi-square test, along with high values of the reliability

Table 2
Summary Statistics

		Raters	Student performances	Items	Domains	Days
Measure (Logits)						
	M	0.26	0.00	0.00	0.00	0.00
	SD	0.10	0.48	0.57	0.28	0.13
Infit *MSE*						
	M	1.00	1.00	1.00	1.02	1.01
	SD	0.05	0.08	0.03	0.03	0.04
Outfit *MSE*						
	M	0.99	0.99	0.99	1.01	1.00
	SD	0.09	0.13	0.06	0.04	0.08
Reliability of separation		0.80	0.91	0.98	0.96	0.92
Chi-square		39.9*	837.1*	822.7*	416.2*	67.3*
df		7	88	27	7	4

Note: *p < 0.01.

of separation statistic for raters (*Rel* = 0.80), student performances (*Rel* = 0.91), items (*Rel* = 0.98), domains (*Rel* = 0.96), and days (*Rel* = 0.92).

Variable map. Figure 2 is a variable map based on Model I that provides a graphical summary of the overall calibration of raters, student performances, items, domains, and days in terms of rater accuracy; this figure corresponds to the summary statistics presented in Table 2. The first column is the logit scale on which the five facets in Model I were calibrated. The second column shows the logit-scale location estimates for the eight operational raters. For raters, higher locations on the logit scale indicate that a rater provided accurate ratings more often, and lower locations on the logit scale indicate that a rater provided inaccurate ratings more often. As can be seen in the figure, Rater 6 was the most accurate rater (location = 0.41 logits, *SE* = 0.05), and Raters 3 and 5 were least accurate (location = 0.11 logits, *SE* = 0.04) across the 5 days.

The third through fifth columns in Figure 2 show the logit-scale location estimates for the individual student performances, items, domains, and days in the solo music performance assessment. For these facets, higher locations indicate that a particular element of the assessment (i.e., an individual performance, item, domain, and day) was easy to score accurately, and lower locations indicate that a particular element of the assessment was difficult to score accurately. As was observed based on the summary statistics for Model I (Table 2), the logit-scale locations shown in the variable map suggest that there was a wide spread in the difficulty associated with assigning accurate ratings across individual performances, items, domains, and days in the assessment procedure. Among the items, the results in Figure 2 suggest that raters were most accurate on Item 14 (location = –1.82 logits, *SE* = 0.17) and least accurate on Item 16 (location = 0.89 logits, *SE* = 0.08) across the 5 days of the music performance assessment. Among the domains, the results in Figure 2 suggest that, overall, the raters were most accurate on Domain E (Visual; location = –0.68 logits, *SE* = 0.03) and least accurate on Domain C (Articulation; location = 0.24 logits, *SE* = 0.08) across the 5 days of the music performance assessment. Because changes in rater accuracy across time are the major focus in this study, the calibration of each of the 5 days in the solo music performance assessment is of particular interest. Accordingly, a detailed discussion of rater accuracy within each of the 5 days is discussed below.

Calibration of the day facet. Table 3 includes logit-scale location estimates for rater accuracy within each day of the music performance assessment along with average differences between rater accuracy within each day and every other day in the assessment procedure. As noted in the discussion of Figure 2, the calibration of the day facet was specified such that higher logit-scale locations indicate more accurate scoring, such that raters provided accurate ratings more often on a particular day, and lower logit-scale locations indicate less-accurate scoring, such that raters provided inaccurate ratings more often on a particular day. In terms of the overall calibration of the 5 days, the location estimates in Table 3 indicate that the raters were most accurate on Day 1 and

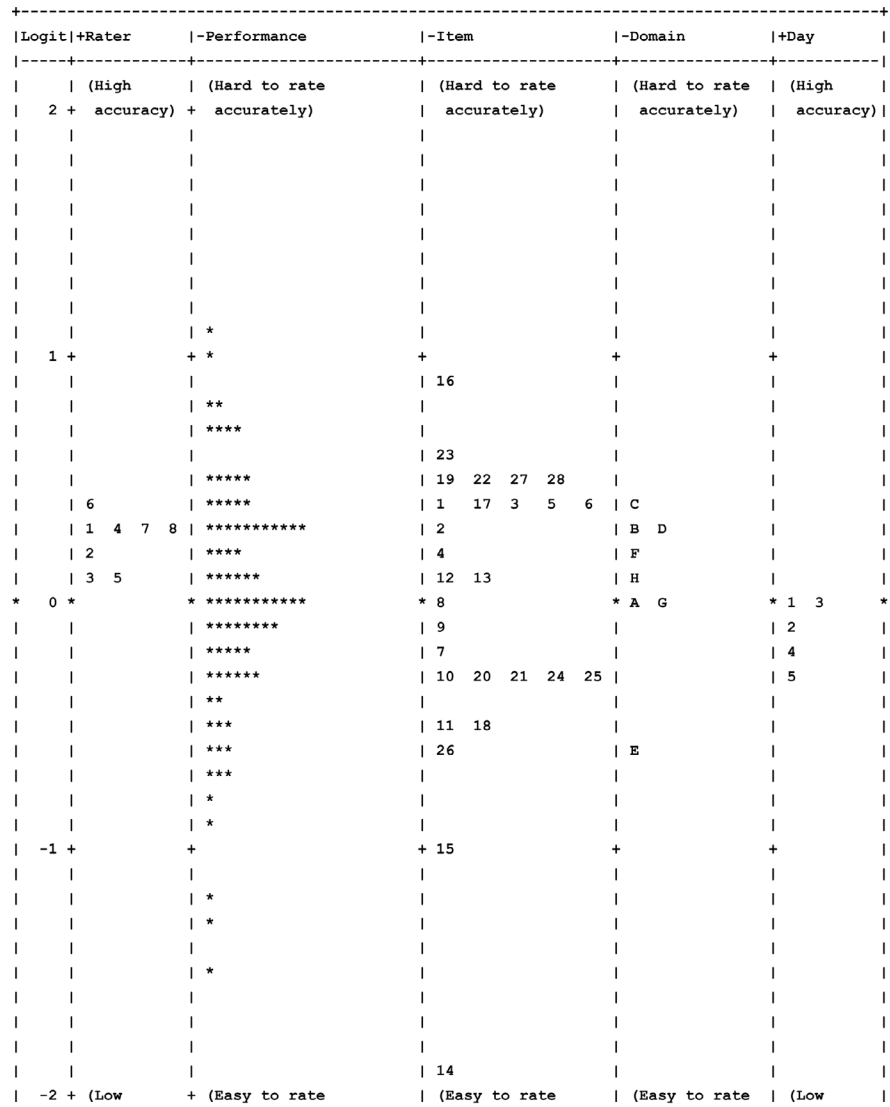

Figure 2. Variable map (Model I).

Table 3
Differences in Rater Accuracy Related to Day

		Mean differences in rater accuracy				
Day	Measure	1	2	3	4	5
1	0.00	—	0.11	0.00	0.22	0.33*
2	–0.11		—	–0.11	0.11	0.22
3	0.00			—	0.22	0.33*
4	–0.22				—	0.11
5	–0.33					—
Chi-square	0.92					
df	67.3*					

Notes: (1) *$p < 0.01$. (2) The logit-scale measures for each day are shown in the second column, and these measures correspond to those illustrated in Figure 1 and summarized in Table 1. (3) The differences shown in the rows of columns 3–7 are calculated as the measure for the day shown in the row minus the measure for the day shown in the column. Accordingly, positive values suggest that raters were more accurate on the day shown in the row compared to the day shown in the column. Negative values suggest that raters were less accurate on the day shown in the row compared to the day shown in the column.

Day 3 (0.00 logits), followed by Day 2 (–0.11 logits), Day 4 (–0.22 logits), and Day 5 (–0.33 logits). Significant differences in logit-scale locations were observed between Day 1 and Day 5 and between Day 3 and Day 5 (Mean difference = 0.33 logits, $p < 0.01$).

Model II

Next, differences in rater accuracy across time were considered at the level of individual raters using an interaction analysis. Overall, the results from the omnibus test for Model II suggested that there was not a significant interaction between the rater and day facets ($\chi^2(40) = 41.3$, $p = 0.41$). This finding suggests that, although there were significant differences in the overall accuracy levels of each individual rater as well as significant differences in overall rater accuracy across the 5 days of the assessment procedure, the relative ordering of the difficulty associated with assigning accurate ratings on each day of the assessment was not significantly different across the eight operational raters.

Model III

Next, differences in rater accuracy within domains were considered across days of the assessment using an interaction analysis. Overall, the results from the omnibus test for Model III suggested that there was a significant interaction between the rater and domain facets ($\chi^2(40) = 193.5$, $p < 0.001$). This finding suggests that the difficulty associated with providing accurate ratings related to each of the domains in the assessment rubric was not consistent across the 5 days of the assessment procedure.

Figure 3 illustrates the results from this interaction analysis in terms of the pairwise combinations of domains and days. Domains are shown along the x-axis, and the value of the *T*-statistic for the interaction between each domain and day is shown on

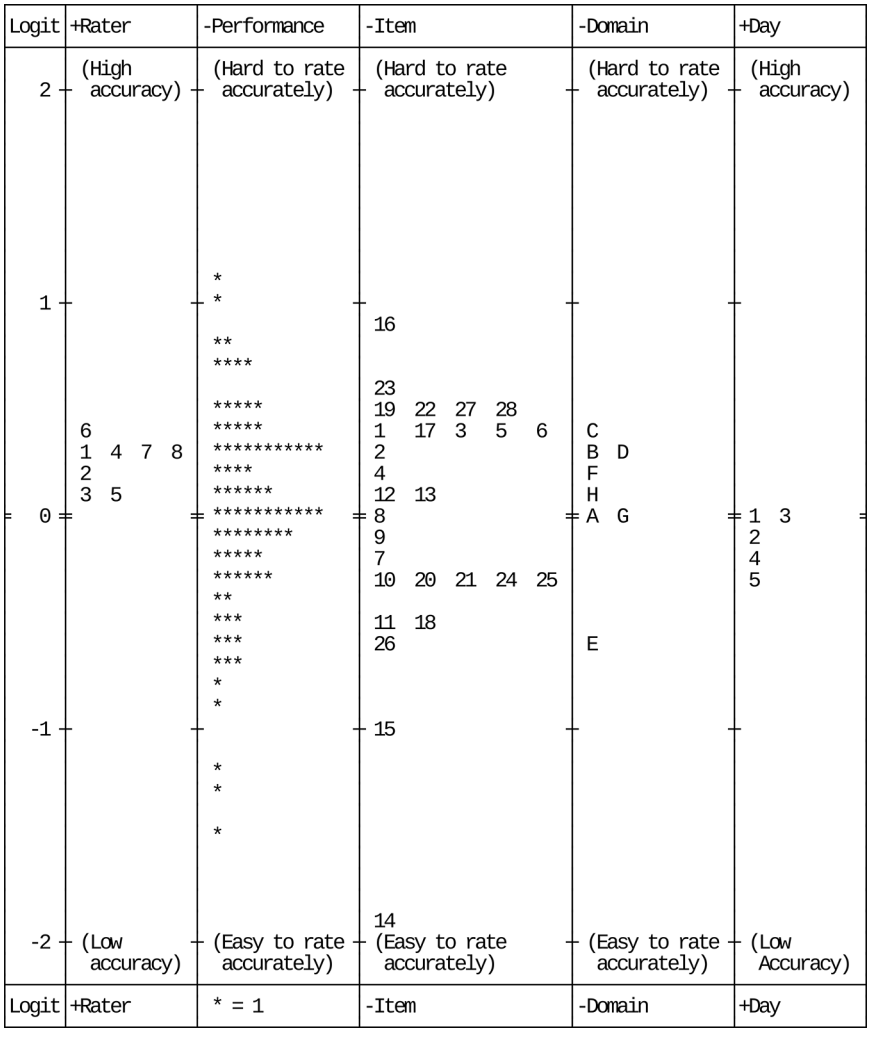

Figure 3. Pairwise interactions for domain * day (Model III).

Note: The y-axis shows the value of the T statistic for the interaction between rater accuracy within each domain (x-axis) and each day of the music solo performance assessment. Numbers are used as plotting symbols to represent the 5 days of the assessment, where Day 1 is represented with "1," Day 2 is represented with "2," and so on. Values of the T statistic that exceed +2 (high dashed line) indicate that raters were significantly more accurate within a domain than expected on a given day, and values of the T statistic below −2 (low dashed line) indicate that raters were significantly less accurate within a domain than expected on a given day.

the y-axis. Numeric plotting symbols are used for each of the 5 days, where the number reflects the day of the assessment. Similar to the results from Model II, values of *T*-statistics that exceed +2.00 indicate that the overall group of raters was significantly more accurate within a particular domain than expected on a particular day, and values below –2.00 indicate that the overall group of raters was significantly less accurate within a particular domain than expected on a particular day. The results from the pairwise interaction analysis illustrated in Figure 3 reflect the overall finding of a significant interaction between the domain and day facets. Specifically, this figure reveals at least one *T*-statistic that falls outside of $-2 \leq T \leq +2$ for each of the domains except Technique, Tone, and Expressive Devices. Furthermore, the magnitude and direction of these significant differences varies across raters and days. For example, the Articulation and Intonation domains only had one significant interaction each that reflected lower-than-expected ratings on Day 4 and Day 5, respectively. For these domains, the interaction results suggest that the difficulty associated with providing an accurate rating increased over the course of the music assessment. In contrast, the Visual, Air Support, and Melody domains had significant interactions in both directions across the 5 days.

It is interesting to note that, for the Visual domain, the significant positive and negative interactions include adjacent days of the assessment, such that is was easier to assign accurate ratings related to visual aspects of a student's performance toward the end of the assessment. However, the opposite pattern occurred for Air Support and Melody, where it was easier to assign accurate ratings related to these components of a student's performance at the beginning of the assessment. A numeric summary of the interaction analysis in terms of individual domains and days that corresponds to Figure 3 is provided in Appendix B here: http://bcrme.press.illinois.edu/media/215/.

SUMMARY AND CONCLUSIONS

The purpose of this study was to explore the degree to which rater accuracy varied across time within the context of a multiday solo music performance assessment. Differences in rater accuracy were considered using criterion-referenced accuracy indices based on Rasch measurement theory. Accuracy scores were calculated by comparing operational raters to an expert rater for each performance, where an exact match between an operational rater and the expert rater resulted in an accuracy score of "1," and any discrepancy resulted in an accuracy score of "0." Using these accuracy scores, three formulations of the MFR-RA model were specified in order to examine differences in rater accuracy as they related to days of the assessment, individual raters, and domains in the scoring rubric. Overall, the results suggested that rater accuracy was not invariant over time for the overall group of raters and that rater accuracy within domains was not invariant over time. These differences in rater accuracy over time have implications for the fairness of the assessment procedure.

In this section, conclusions are presented as they relate to the three research questions for this study. A discussion of the implications of these findings in terms of research and practice follows.

Are There Overall Differences in Rater Accuracy Across the Duration of the Music Solo Performance Assessment?

In order to evaluate the degree to which rater accuracy varied across the 5 days of the music performance assessment, a MFR-RA model was specified that included a facet for days of the assessment procedure. Using this model, it was possible to estimate the location of each of the 5 days on a linear scale (the logit scale) that represented rater accuracy. The results revealed differences in accuracy for the overall group of raters over the 5 days of the music solo performance assessment. Specifically, the overall group of raters was significantly more accurate on the first day of the assessment compared to the last day. A significant difference was also observed between the third day of the assessment and the last day of the assessment, where raters were more accurate on the third day than the last day. Together, these results suggest that that rater accuracy decreased over the course of the music assessment.

Do Any Individual Raters Demonstrate Interactions Between Accuracy and Days of the Assessment?

The second research question focused on the degree to which the difficulty associated with providing accurate ratings across the 5 days of the music assessment was consistent across the individual raters. In order to explore this research question, a MFR-RA model was specified that included an interaction term between days and individual raters (Model II). Results from the omnibus test for the interaction term suggested that there were not significant interactions between days of the music assessment and individual raters in terms of accuracy. Accordingly, it is possible to interpret the overall ordering of rater accuracy within the 5 days of the assessment procedure consistently across the individual raters. In other words, the changes observed in rater accuracy over time appeared to affect all of the raters in a similar fashion.

Is Rater Accuracy Within Domains Invariant Across Days of the Assessment?

The third research question focused on the degree to which rater accuracy within domains was consistent across the 5 days of the music assessment. In order to explore this research question, a MFR-RA model was specified that included an interaction term between domains and days (Model III). Results from the omnibus test for the interaction term suggested that rater accuracy within domains was not invariant across the days of the music assessment. This finding suggests that the difficulty associated with providing accurate ratings related to different aspects of solo music performances was not consistent over the course of the 5-day assessment. Pairwise interaction analyses between individual domains and each of the days in the assessment revealed significant differences in rater accuracy within the Articulation, Intonation, Visual, Air Support, and Melody domains across the days of the assessment. With the exception of the Visual

domain, all of the significant interactions indicated lower-than-expected rater accuracy on the last days of the assessment and higher-than-expected rater accuracy during the first days of the assessment—suggesting that the difficulty associated with providing accurate ratings in these domains increased over time. In contrast, the significant interactions related to the Visual domain suggested that this component of student performances became easier to score accurately over time.

DISCUSSION

This study contributes to previous research related to the examination of rating quality in performance assessment in general as well as within the context of music performance assessment more specifically. In terms of performance assessment research in general, several researchers have considered rating quality from a dynamic perspective (Harik et al., 2009; Hoskens & Wilson, 2001; Leckie & Baird, 2011; Wolfe et al., 2001). However, this work has not emphasized changes in rater accuracy within a criterion-referenced framework based on expert raters. In terms of music assessment research, this study was also the first of its kind to evaluate accuracy indices in conjunction with rater errors from a dynamic perspective.

The findings from this study have several important implications for research and practice related to music performance assessment. In terms of research, the results from this study highlight the dynamic nature of rating quality over the course of multiday solo music performance assessments. Although the overall accuracy calibrations for raters suggested that the raters were generally accurate, the results indicated that the overall group of raters became less accurate over time and that these changes in rater accuracy affected domains in the assessment rubric differently over the course of the scoring period. As a result, researchers who are interested in exploring rating quality in music assessments should consider the potential influence of time on the quality of rater judgments. From a methodological perspective, this study builds upon previous work related to changes in rating quality across time in the context of a music performance assessment (Wesolowski, Wind, & Engelhard, in press) to include a procedure for considering rater accuracy (Wesolowski & Wind, 2017) from a dynamic perspective. Specifically, researchers can apply the method illustrated in this study for evaluating rater accuracy from a dynamic perspective to other multiday music performance assessments in order to empirically evaluate the degree to which rater accuracy is invariant over time. In particular, an indication of variability due to rater accuracy over time may provide more grounded evidence for the validity and fairness arguments of multiday assessment contexts. As with any other evaluation of psychometric quality, evidence of changes in rater accuracy over time should be interpreted in light of the unique characteristics of each assessment context.

In terms of practice, the findings of this study offer two important implications. First, this study provides a new methodology in music for evaluating the fairness of

music performance assessments. The ability to identify the phenomenon of raters' differential accuracy as a dynamic process while also simultaneously collecting concrete evidence of the specific manner in which the raters apply the measurement instrument provides a clear and understandable mechanism for reporting and improving upon the fairness of the music performance assessment. Because fairness is fundamentally a validity issue, the reporting of raters' differential accuracy over the course of a multiday music performance assessment can provide important validity evidence of the inferences drawn from the assessment regarding student and or ensemble performance achievement. This is exceptionally important in today's data- and accountability-driven educational environment when student achievement data is linked with teacher effectiveness. More so, evidence of rater behavior is important for the validity of an assessment context, particularly when the National Association for Music Education's "Position Statement on Assessment in Music Education" (2017) clearly indicates that teachers should "be certain to include the outcomes of traditional festival rankings, as these are one legitimate tool for assessing the quality of school music programs."

Second, the findings of this study provide a foundational methodology to be used in practice in rater training protocols specific to the field of music performance beyond the specific assessment described in the current study. Our study demonstrated a new application of methods for detecting rater DRIFT within the context of music performance assessment, particularly by combining DRIFT indices (Wesolowski, Wind, & Engelhard, in press) with accuracy indices (Wesolowski & Wind, 2017). Several rating quality indices previously introduced by the authors in the research literature included static evaluation of raters' leniency/severity (Wesolowski et al., 2016a), precision in raters' use of a measurement instrument's rating scale structure (Wesolowski et al., 2016b), raters' systematic differential severity based upon performers' subgroup affiliation (Wesolowski et al., 2015), and raters' differential severity of item use due to personal idiosyncrasies defined by rater type (Wesolowski, 2017). Each of these indices provided valuable information regarding particular aspects of rater behavior. However, these indices, arguably, provide diagnostic evidence of rater behavior but not formative evidence; this means the phenomenon of differential rater severity/leniency over time can be detected, but there is no evidence of where the variability occurs with the interaction between the rater and the measurement instrument. Because rater DRIFT indices combined with accuracy indices provide tangible verification of the raters' specific use of the measurement instrument, the methodology of combined indices allows for the ability to directly isolate domains and items within the measurement instrument that demonstrate variability throughout the course of the assessment. Isolation of problematic areas is important, as it allows the ability to quickly recalibrate raters on problematic areas and provide real-time feedback of each rater's behavior and overall rating quality. As a result, if rater behavior were to be monitored in real time throughout other multiday music performance assessments that are similar to this study, raters could be recalibrated when the variability is discovered, thereby improving the fairness of the remainder of the assessment context.

Actionable Steps for Music Assessments

In order for these procedures to be implemented effectively, we suggest three broad actionable steps in establishing quality control of raters and implementation of related data analysis procedures in the context of formal music performance assessment: (1) rethinking of rater training protocols, (2) rethinking of rater selection protocols, and (3) addition of data analysis protocols. First, there is a large body of research on the effectiveness of empirical rater training methods to monitor rater quality in educational performance assessment contexts (see Raczynski, Cohen, Engelhard, & Lu, 2015). We suggest the investigation into the effectiveness of developing and maintaining high-quality raters in the context of formal music performance assessment. Second, we suggest using rater-training protocols as a mechanism for selecting raters. With the understanding that even highly experienced subject matter experts demonstrate variability in the scoring process, we recommend that the process of selecting raters for formal music assessments should shift from experience-based selections to empirical-based selections, specifically where selection criteria are based upon a prospective rater's ability to be trained and to provide high-quality, accurate ratings according to an established standard and/or expectation. Third, we suggest the planned use of psychometric (e.g., rater quality) analyses throughout the assessment context. In order to implement real-time rater feedback, we encourage the use of psychometric experts to collect, calibrate, analyze, monitor, and report adjudicator ratings in real time or at the end of the daily evaluations using the methodologies described in this study. The analysis can then be used to inform raters of their behavior with specific feedback to address and change problems and inconsistencies with the scoring data. It is important to note that in each of these three steps, empirical evidence of rating quality is the foundation for decision-making.

LIMITATIONS

When considering the results from this study, it is important to note a potential limitation related to the conceptualization of rater accuracy based on a single expert's ratings. Because a single expert's judgments of the musical performances was used as the criteria against which the operational raters were evaluated, any construct-irrelevant idiosyncrasies in the expert's judgment would compromise the interpretation of the accuracy measures. However, as noted above, the experiences of the expert rater, the expert's direct involvement in the development of the scoring materials, and empirical evidence to support the psychometric quality of the expert's ratings (see Wesolowski, Wind, & Engelhard, in press) establish the expert's judgments as a meaningful criterion against which the operational raters' judgments could be evaluated.

FUTURE DIRECTIONS AND RECOMMENDATIONS

Moving forward, we recommend the construction of rater-training protocols and rater-monitoring protocols for music performance assessments. It is clear that through the

variety of studies cited within this article, rater variability poses a serious threat to the validity and fairness of formal music performance assessments both from static and dynamic perspectives. Training for raters prior to a music performance assessment should include practice with scoring on the measurement instrument and particular access to anchor recordings representing each of the levels of performance by domain and item. We recommend that raters must achieve a threshold of accuracy scores, designated by those in charge of the assessment, prior to the admittance of judging formal music performance assessments. It is clear that even the most highly qualified of content experts introduce construct-irrelevant variability into the assessment context, thereby affecting the validity and fairness of the assessment. Qualifications such as teaching experience, initial 1-day trainings, documented success in teaching, adjudicator experience, and music degree requirements simply are not enough to qualify raters for valid and fair formal music performance assessments (Texas, n.d.; Florida, n.d.; Ohio, n.d.). Furthermore, we recommend the development of rater-monitoring protocols, where the quality of raters is monitored in real time throughout the course of multiple day assessments based on both observed ratings and rater accuracy indices that incorporate expert ratings. This monitoring can be implemented in operational music assessments using processes similar to "read-behinds" in the context of writing assessment, where expert raters rate a subset of performances that have been scored by operational raters, and expert ratings are compared to operational ratings (Hoskens & Wilson, 2001; Knoch, Read, & von Randow, 2007). Alternatively, benchmark performances with established ratings (e.g., from expert raters) can be interspersed among the set of performances during the course of the performance assessment in order to gauge rater accuracy over time based on the match between operational and expert ratings on the benchmarks (Wang, Song, Wang, & Wolfe, 2017).

SUPPLEMENTAL MATERIAL

Appendixes A and B are available online at https://bcrme.press.illinois.edu/media/215/.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

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.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56–70.

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

Harik, P., Clauser, B., Grabovsky, I., Nungester, R., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58.

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*(2), 121–145.

Kane, M. T. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418.

Linacre, J. M. (2015). Facets Rasch Measurement (Version 3.71.4). Chicago, IL: Winsteps.com.

Myford, C., & 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.

National Association for Music Education. (2017). *Assessment in music education (position statement).* Retrieved from http://www.nafme.org/about/position-statements/assessment-in-music-education-position-statement/assessment-in-music-education/

Raczynski, K., Cohen, A., Engelhard, G., Jr., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement, 52*(3), 301–318. https://doi.org/10.1111/jedm.12079

Union, I. T. (2004). *Objective perceptual assessment of video quality: Full reference television*. *ITU-T Telecommunication Standardization Bureau*. Geneva, Switzerland.

Wang, C., Song, T., Wang, Z., & Wolfe, E. (2017). Essay selection methods for adaptive rater monitoring. *Applied Psychological Measurement, 41*(1), 60–79. https://doi.org/10.1177/0146621616672855

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., Athanas, M., Burton, J., Edwards, A. S., Edwards, K. E., Goins, Q., . . . Thompson, J. (in press). Judgmental standards setting: The development of objective content and performance standards for secondary-level solo instrumental music assessment. *Journal of Research in Music Education.*

Wesolowski, B. C., & Wind, S. A. (in press). Validity, reliability, and fairness. In T. Brophy (Ed.), *The Oxford handbook of assessment policy & practice in music education*. New York, NY: Oxford University Press.

Wesolowski, B. C., & Wind, S. A. (2017). Investigating rater accuracy in the context of secondary-level solo instrumental music performance. *Musicae Scientiae*. Advance online publication. doi:10.1177/1029864917713805

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

Wesolowski, B., Wind, S. A., & Engelhard, G., Jr. (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., Jr., (2016a). Quality control of rater analyses in music performance assessment: Application of the many facet Rasch model. In T. Brophy (Ed.), *Connecting practice, measurement, and evaluation: The Fifth International Symposium on Assessment in Music Education* (pp. 335–356). Chicago, IL: GIA Publications.

Wesolowski, B., Wind, S. A., & Engelhard, G., Jr. (2016b). 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.

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

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

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, G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition Examination using benchmark essays* (Research Report No. 2007–2). New York, NY: College Entrance Examination Board.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10. https://doi.org/10.1016/j.asw.2015.06.002