Article

MUSICÆ SCIENTIÆ

Musicae Scientiae
1–20

DOI: 10.1177/1029864917713805
journals.sagepub.com/home/msx
SAGE

Investigating rater accuracy in the context of secondary-level solo instrumental music performance

Brian C. Wesolowski
The University of Georgia, USA

Stefanie A. Wind
The University of Alabama, USA

Abstract

In any performance-based musical assessment context, construct-irrelevant variability attributed to raters is a cause of concern when constructing a validity argument. Therefore, evidence of rater quality is a necessary criterion for psychometrically sound (i.e., valid, reliable, and fair) rater-mediated music performance assessments. Rater accuracy is a type of rater quality index that measures the distance between raters' operational ratings and an expert's criterion ratings on a set of benchmark, exemplar, or anchor musical performances. The purpose of this study was to examine the quality of ratings in the context of a secondary-level solo music performance assessment using a Multifaceted Rasch Rater Accuracy (MFR-RA) measurement model. This study was guided by the following research questions: (a) overall, how accurate were the rater judgments in the assessment context? (b) how accurate were the rater judgments across each of the items of the rubric?, and (c) how accurate were the rater judgments across each of the domains of the rubric? Results indicated that accuracy scores generally matched the expectations of the MFR-RA model, with rater locations higher than the average student performance, item, and domain locations, indicating that the student performances, items, and domains were relatively easy to rate accurately for the sample of raters examined in this study. Overall, rater accuracy ranged from 0.54 logits (SE = 0.05) for the most accurate rater to 0.24 logits (SE = 0.04) for the least accurate rater. Difficulty of rater accuracy across items indicated a range of 0.91 logits (SE = 0.08) to -1.83 logits (SE = 0.17). Difficulty of rater accuracy across domains ranged from 0.25 logits (SE = 0.08) to -0.68 logits (SE = 0.17). Implications for the improvement of music performance assessments with specific regard to rater training are discussed.

Keywords

accuracy, assessment, music performance, Rasch measurement, raters, rubric

Corresponding author:
Brian C. Wesolowski, Hugh Hodgson School of Music, The University of Georgia, 250 River Road, Athens, GA 30602, USA.
Email: bwes@uga.edu

The National Association for Music Education (NAfME), on its broad concerns of assessment in the field of music education within the United States, outlined several guidelines for conducting effective assessments in its *Assessment in Music Education Position Statement* (2017). Some guidelines include processes for selecting quality assessment tools, a focus on developing uniform assessments across specified disciplines, and working to include results of musical assessments within more broad educational reporting mechanisms. In terms of the value of including formal, festival-style music performance assessment ratings in teachers' reporting processes, the position statement specifically states: "Be certain to include the outcomes of traditional festival rankings, as these are one legitimate tool for assessing the quality of school music programs" (para. 9). The concern of the legitimacy of these assessment contexts, however, was brought to light by Hash (2013). One important consideration was related to the skew of ratings as evidenced from several rating data sources. From a descriptive analysis perspective, he provided evidence of multiple studies that indicate an overall negative skew of ensemble ratings based upon a typical five-category structure (e.g., superior, excellent, good, fair, poor). Hash indicated, "It is unclear why ratings trend toward the top of the scale" (p. 165).

From a qualitative perspective, only a phenomenographic investigation into each individual rater can provide concrete evidence of why he or she may have assigned particular ratings to ensembles in a particular manner (Wesolowski et al., 2017; Wesolowski, Burrack, & Parkes, in press). Any possible considerations outside of a direct investigation into each individual rater are only speculative at best. From a quantitative perspective, measurement error of any music performance assessment context can be attributed to four specific factors: (a) the ability of the performer or ensemble, (b) the difficulty of the task, (c) variability in rater judgments, and (d) the manner in which the rater applies the measurement instrument (Wesolowski, Wind, & Engelhard, 2016b). This study specifically investigated the last factor: the manner in which the rater applies the measurement instrument.

The role of the rater in direct musical assessments

A direct music performance assessment can be defined as a type of musical assessment where a sample of an individual person's or ensemble's musical ability is obtained under controlled testing conditions and then evaluated either concurrently or subsequently by one or more expert music content raters. Direct music performance assessments are commonly used as the primary type of assessment structure in formal music evaluation procedures such as solo and ensemble evaluations, all-state auditions, and large-ensemble performance assessments, as the act of evaluating students "doing" music provides a strong face validity argument to the assessment context. More specifically, direct music performance assessments represent and most closely simulate the actual teaching and learning in the music performance classroom.

From a testing perspective, a direct music performance assessment is based upon a constructed-response format, where observed scores are generated from raters' interpretations of a set of evaluation cues (e.g., domains, items, and rating scale categories) provided within a measurement instrument. Although direct assessments provide strong evidence of face validity, the crafting of more robust validity arguments for constructed-response assessment formats is challenging because the reliance on raters to mediate the assessment process can introduce unwanted construct-irrelevant variance into the assessment process. According to Eckes (2012), rater variability in the context of any constructed-response assessment format 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. Empirical evidence of construct-irrelevant variability stemming from raters is perhaps no surprise, as any psychical activity is based upon rapid and subjective decision-making (Freud, 1920). Based upon Hash's (2013) observations, it is clear that in the context of formal performance assessments, the field of music has succumbed to interpreting the results as directly associated with the degree to which the rater demonstrates qualities of leniency or severity in their rating outcomes (Shavelson & Webb, 1991, p. 8). As Bond and Fox (2015) noted: "if we were serious about measurement of ability in a judged situation, we would want to be sure that passing or failing depended more on candidate ability than on luck of the draw with examiners" (p. 169). As a result of raters' personal idiosyncrasies, a thorough investigation into rater quality is one of the most important considerations for building validity arguments in the context of any direct music performance assessment.

In music assessment research, the examination of rater behavior can be classified into two distinct categories: behavior-centered approaches and empirical-centered approaches (see Wesolowski et al., 2016b, for a thorough examination). Behavior-centered approaches investigate the ecological content of rater judgment, including attributes of the raters themselves (e.g., experiences, content knowledge, cognition, perception, and thought processes) and/or their surrounding environment (e.g., assessment condition and performer-effects). The second category is an empirical-centered approach. This approach investigates rater quality and rater effects using statistical indices that provide evidence of the overall quality of assigned ratings during the operational scoring procedures of a musical assessment. As noted by Murphy and Cleveland (1991) and Johnson, Penny, and Gordon (2009), evidence of rater quality can be demonstrated through three distinct indices: (a) rater agreement, (b) rater errors and systematic biases, and (c) rater accuracy. Wind and Engelhard (2013) describe these three indices as follows:

> Indices of rater agreement describe the degree to which raters assign matching scores to the same performance; these indices include measures of categorical agreement and association among raters. Rater errors and systematic biases are used to describe specific patterns or trends in rating behavior that are believed to contribute to the assignment of scores different from those warranted by a student's performance. Rater accuracy is defined in practice as a match between operational ratings and those established as "true" or "known" ratings by individuals or committees of expert raters (p. 280).

Although these indices are all different from a statistical perspective and address slightly different research questions, all three categories of indices similarly attempt to identify systematic rating patterns that may contribute unwanted construct-irrelevant variance into the assessment context, thereby improving its overall quality.

As the field of music is becoming more familiar with the philosophical, theoretical, and applied applications of Item Response Theory (ITR), its use in the process of examining rater behavior is becoming increasingly used to build validity arguments for various music performance assessments (Wesolowski, 2017). In particular, Rasch Measurement Theory is becoming increasingly pervasive in the field of music as a method for providing evidence of measurement invariance of persons, items, raters, and other facets of interest. Rasch (1960) developed a set of probabilistic measurement models that specify the requirements of invariant measurement. The concept of rater invariant measurement is simple in the context of direct music performance assessments: the measurement of two performers with theoretically equal musical ability should be evaluated identically without any effect of the rater. In the case of direct performance-based assessments where raters mediate the assessment process, five

requirements of rater-invariant measurement can be used as a method to test the hypothesis of invariant measurement with specific attention to raters: (a) rater invariant measurement of persons (i.e., the measurement of performers must be independent of the particular raters that happen to be used for the measuring); (b) non-crossing person response functions (i.e., a higher achieving performer must always have a better chance of obtaining higher ratings from raters than a lower achieving performer); (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular performers used for calibration); (d) non-crossing rater response functions (i.e., any performer must have a better chance of obtaining a higher rating from lenient raters than from more severe raters); and (e) variable map (i.e., performers and raters must be simultaneously located on a single underlying latent variable) (adapted from Engelhard, 2013).

In the context of secondary-level music performance assessment, recent empirical-centered investigations into rater quality (i.e., rater agreement, rater errors, and systematic biases) using Rasch measurement theory demonstrate that even the most qualified of content experts demonstrate construct-irrelevant variability due to rater errors. Examples of rater variability in music research literature include marked evidence of raters' leniency/severity (Wesolowski, Wind, & Engelhard, 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, Wind, & Engelhard, 2015), raters' differential severity of item use due to personal idiosyncrasies defined by rater-type (Wesolowski, 2017), the time-ordering in which raters evaluate a performance within a particular assessment context (Wesolowski, Wind, & Engelhard, in press), and the placement of a rater within a specified rater linking design (Wind, Engelhard, & Wesolowski, 2016). These studies each provide empirical evidence of variability in raters' judgments based upon the principles of rater-invariant measurement; however, these investigations do not provide direct evidence for the manner in which the rater applies the measurement instrument.

Evidence for the manner in which the rater applies the measurement instrument is demonstrated through evidence of rater accuracy. Rater accuracy is defined as "the match between ratings obtained from operational raters and those obtained from an expert … on a set of benchmark, exemplar, or anchor performances" (Engelhard, 1996, p. 56). Engelhard's definition implies that accuracy is a type of rater quality index that measures the distance between operational ratings and criterion ratings. The criterion ratings, in the case of this study, are the ratings defined by the expert rater. Engelhard (2013) further elaborates that "accuracy can be defined as the comparison between unknown processes and a defined standard process in order to adjust the unknown process until it matches the standard" (p. 232). As opposed to the rater quality indices indicated previously (e.g., rater agreement, rater errors, and systematic biases), the advantage of rater accuracy indices is that they provide specific and concrete evidence in the ways a rater applies a measurement instrument to an assessment context against a defined standard of rating.

Rater accuracy data provides important diagnostic information that allows for not only the ability to calculate distances between operational ratings and the accepted standard, but it also allows for the ability to retrain or recalibrate a rater when evidence of rater errors are found within the assessment context. As Murphy and Balzer (1989) point out, rater quality indices such as rater agreement, rater errors, and systematic biases provide indirect measures of rater accuracy and are important rater quality indices when direct measures of rater accuracy are not available. However, the addition of rater accuracy indices to an assessment context, when available, provide a more detailed analysis of rater quality when in conjunction with rater agreement, rater errors, and systematic biases indices, thereby providing more thorough validity evidence of a direct music performance assessment.

Rater accuracy methodology

A variety of techniques for evaluating rater accuracy appear throughout the psychometric literature on rating quality that reflect two broad conceptualizations of accuracy in the context of rater-mediated assessments: (a) agreement accuracy; and (b) criterion-referenced accuracy. Essentially, the *agreement accuracy* perspective reflects a view of high levels of rater agreement or consistency in examinee ordering across raters (i.e., rater reliability) as evidence of accuracy. When this perspective is applied in rating quality analyses, evidence of alignment between individual raters and the overall group of raters is used as evidence of rater accuracy. Common accuracy indicators that reflect this perspective include reliability and generalizability coefficients based on Generalizability Theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), where higher levels of reliability and generalizability are viewed as evidence of rater accuracy.

On the other hand, the *criterion-referenced accuracy* perspective reflects a view of alignment between observed ratings and criterion ratings as evidence of accuracy (Wesolowski, Wind, et al., in press). Criterion ratings reflect "known" or "true" ratings for student performances that provide a frame of reference for evaluating operational raters. In operational rater-mediated assessment systems, criterion-referenced accuracy indicators are implemented through the use of "read-behind" procedures (Hoskens & Wilson, 2001; Johnson et al., 2009; Knoch, Read, & von Randow, 2007), where expert raters rate subsets of common performances with operational raters, and compare their ratings with those of operational raters. Other practical applications of this approach include interspersing performances with criterion ratings among the set of performances scored by operational raters in order to evaluate the alignment between operational and criterion ratings in an ongoing fashion (e.g., Wang, Song, Wang, & Wolfe, 2017).

Within the framework of Classical Test Theory (CTT), criterion-referenced accuracy indicators include agreement statistics between operational and criterion ratings (Johnson et al., 2009). On the other hand, Item Response Theory (IRT) indicators of criterion-referenced accuracy include accuracy models where the match between the ratings that operational and criterion raters assign to the same performance is treated as a latent variable on which individual raters, student performances, and other facets are calibrated. This approach was originally presented by Engelhard (1996) within the framework of Rasch measurement theory. The match between operational and criterion ratings can be defined using a dichotomous scoring scheme, where a critical value establishes the maximum difference between operational and criterion ratings that can be classified as accurate, such that ratings with a larger discrepancy are classified as inaccurate. Table 1 summarizes Engelhard's indices of rater accuracy based upon the Many-Facet Rasch (MFR) model used in this study.

Multifaceted Rasch rater accuracy measurement model

Following the dichotomous formulation presented by Engelhard (1996), the Rasch rater accuracy model is stated mathematically as:

$$\ln\left[\frac{P_{ij(x=1)}}{P_{ij(x=0)}}\right] = \lambda_i - \beta_j \tag{1}$$

where

$P_{ij(x=1)}/P_{ij(x=0)}$ = the probability that rater i provides an accurate rating ($x = 1$), rather than an inaccurate performance ($x = 0$) on student performance j;

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	1. What is the accuracy location of each rater?	1a. Variable map 1b. Calibration and location of elements within facet
	2. Rater accuracy precision	2. How precisely has each rater been calibrated in terms of accuracy?	2. Standard errors for raters
	3. Rater accuracy separation	3.1 How spread out are the individual raters in terms of accuracy?	3.1. Reliability of separation statistic for raters
		3.2 Are the raters considered to be exchangeable in terms of accuracy?	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 MSE)

Note. Adapted from Wind & Engelhard (2013).

λ_i = the ability of rater *i* to provide accurate ratings; and
β_j = the difficulty associated with providing an accurate rating to student performance *j*.

Using a Many-Facet Rasch (MFR) model approach, other facets can be added to Equation 1 whose locations reflect the overall accuracy associated with aspects of the assessment system, such as the difficulty of assigning accurate ratings on the domains in an analytic rubric (Wind & Engelhard, 2013) or to performances with certain characteristics (Wolfe, Song, & Jiao, 2016). The extensions used in this study are elaborated below in the Data Analysis section.

Purpose and research questions

The purpose of this study was to examine the quality of operational ratings in the context of secondary-level solo music performance assessment using a Multifaceted Rasch Rater Accuracy (MFR-RA) measurement model. The research questions that guided this study include:

1. Overall, how accurate were the rater judgments?
2. How accurate were rater judgments across each of the items of the rubric?
3. How accurate were rater judgments across each of the domains of the rubric?

Method

Instrument

The measurement instrument used in this study was the Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO, see Appendix 1) (Wesolowski et al., 2017). The rubric was first developed and validated as a 30-item rubric using the Multifaceted Rasch Partial Credit measurement model. It was developed with the aid of 13 music subject

matter experts (e.g., secondary-level music educators) over the course of a six-week process. The instrument was subjected to a secondary validation study where it was employed to develop content and performance standards using a modified Bookmark Standard Setting Procedure (Wesolowski et al., under review). The measurement model in this secondary study was also the Multifaceted Rasch Partial Credit measurement model. The judgmental standard setting procedure used a panel of eight music context experts different from the original development study. The second validation process provided empirical evidence that two of the original items were slightly overfit, therefore they were removed for this study. As a result, the current operational structure of the rubric consisted of a total of 28 items, each ranging from two to four rating scale category structures, 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).

Performance stimuli

A total of 88 secondary-level (e.g., grades 6–12) video performances were collected and 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 range in performance quality was representative of secondary-level (grades 6–12) solo music performances. The average years of performing experience for the students participating in the study were 3.21 (SD = 1.91). The sample of performing students was gathered from two suburban schools (high school, n = 1; middle school, n =1). There were no requirements for the piece of music chosen. Students were asked to perform any piece of written music they felt comfortable performing. All performances were recorded with the student in a seated position, while reading their selected piece of music on a black Manhasset music stand. The range of performance times of the videos performances was from 33 seconds to 2:35 minutes. All performances were recorded in front of a grey photographer's canvas in order to eliminate visual distractions. No students were directly affiliated with any researchers in this study, and provided informed consent forms along with their legal guardian as approved by the primary investigator's university-affiliated ethics review board.

Participants

Ratings for this rater accuracy analysis came from eight operational raters (raters 1–8) and one expert rater (rater 9). The operational raters were defined as subject matter experts (SMEs) 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 SMEs were active secondary-level (i.e., grades 6–12) instrumental music teachers while concurrently participating in this study. The eight SMEs 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), highest degree earned (bachelors, n = 4; masters, n = 4), and primary instrument (woodwind, n = 4; brass, n = 4). The expert rater has expertise in the research area of measurement and evaluation, and also played a fundamental role in the development of the MPR-2L-INSTSOLO. The expert rater, however, was not part of this present study. Because this rater was directly involved in the development of this instrument, the individual was considered an expert in the construct defined by the rubric. Furthermore, the rater was deemed appropriate based upon their demonstrated moderate severity and adequate fit statistics based upon analysis of the

observed ratings (see Results section below). Accordingly, this rater's judgments can be viewed as a meaningful criterion against which to compare operational rater judgments that is congruent with the selection of expert raters in previous applications of the MFR-RA model (e.g., Engelhard, 1996; Wind & Engelhard, 2013; Wolfe et al., 2016). Each of the nine raters evaluated all 88 video performances using the MPR-2L-INSTSOLO. Performances were provided in randomized order from the video pool. Ratings were collected through an online rubric platform with data stored and collected in a Google Docs spreadsheet.

Data analysis

The data analysis procedure involved two major steps. First, it was necessary to calculate rater accuracy scores based on the match between operational raters and the expert rater for each performance. In keeping with the approach used in the original presentation of the Rasch rater accuracy model (Engelhard, 1996), a dichotomous scoring scheme was applied. For this study, an accuracy score of "1" was assigned when there was an exact match between an operational rater and the expert rater, and a score of "0" was assigned otherwise.

Second, rater accuracy was explored using a MFR model formulation of the Rasch rater accuracy model. This formulation was selected in order to systematically explore differences in rater accuracy across raters, student performances, items, and domains. Specifically, the following formulation of the rater accuracy model was applied using the *FACETS* computer program (Linacre, 2015):

$$\ln\left[\frac{P_{ijm(x=1)}}{P_{ijmk(x=0)}}\right] = \lambda_i - \beta_j - \delta_m - \eta_k, \tag{2}$$

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.

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.

The MFR model in Equation 2 was used to obtain location estimates for each rater, student performance, item, and domain on a linear scale that reflects rater accuracy. Differences among the logit-scale locations for individual raters, student performances, items, and domains reflect differences in levels of accuracy among elements within each of these facets.

Results

Summary statistics

Table 2 includes average logit-scale locations, model-data fit statistics, and separation statistics for each of the facets in the MFR rater accuracy model used in this study.

In order to provide a frame of reference for interpreting the logit-scale locations within each facet, the average locations were centered at zero logits for all of the facets except raters.

Table 2. Summary statistics.

	Raters	Student performances	Items	Domains
Measure (Logits)				
M	0.38	0.00	0.00	0.00
SD	0.10	0.38	0.56	0.29
Infit *MSE*				
M	1.00	1.00	1.00	1.01
SD	0.05	0.09	0.03	0.02
Outfit *MSE*				
M	0.99	0.99	0.99	1.01
SD	0.09	0.12	0.05	0.04
Reliability of separation	0.80	0.85	0.98	0.96
Chi square	39.2*	427.6*	820.4*	431.4*
df	7	87	27	7

Note. *$p < .01$.

Examination of the logit scale locations reveal that, on average, rater locations were higher than the average student performance, item, and domain locations ($M = 0.38$; $SD = 0.10$). This finding suggests that the student performances, items, and domains were relatively easy to rate accurately for the sample of raters examined in this study. In terms of model-data fit statistics, average values of the Infit and Outfit *MSE* statistics suggest that the accuracy scores generally matched the expectations of the MFR model for rater accuracy, with average values around 1.00 across all facets. This finding suggests that estimates of rater, student performance, item, and domain locations on the logit scale can be interpreted as indicators of rater accuracy, and the difficulty associated with providing accurate ratings across individual student performances, items, and domains, respectively.

Finally, the reliability of separation (*Rel*) statistics and Chi Square (χ^2) statistics for each facet suggest that there were significant differences among the logit-scale locations of individual raters ($Rel = 0.80$; $\chi^2(7) = 39.2$, $p < .001$); individual student performances ($Rel = 0.85$; $\chi^2(87) = 472.6$, $p < .001$), individual items ($Rel = 0.98$; $\chi^2(27)=820.4$, $p < .001$), and individual domains ($Rel = 0.96$; $\chi^2(7) = 431.4$, $p < .001$). Together, these results indicate that there are meaningful differences in the difficulty associated with providing accurate ratings across student performances, items, and domains. These differences are explored further below.

Variable map. Figure 1 is a variable map that provides a graphical summary of the results from the MFR rater accuracy model.

The first column is the logit scale on which each rater, student performance, item, and domain was calibrated in terms of the rater accuracy construct. The second column shows rater calibrations. For raters, higher locations indicate higher overall accuracy, and lower locations indicate lower overall accuracy. Accordingly, the results in Figure 1 suggest that Rater 6 was most accurate, and Raters 3 and 5 were least accurate. The next three columns show logit-scale locations for individual student performances, items, and domains. Across these three facets, higher locations suggest that a particular element was difficult to rate accurately, and lower locations suggest that a particular element was easy to rate accurately. Student performance locations are illustrated using an asterisk (*), where an asterisk represents two performances. The item ID numbers and domain numbers were used to illustrate item and domain locations on the logit-scale.

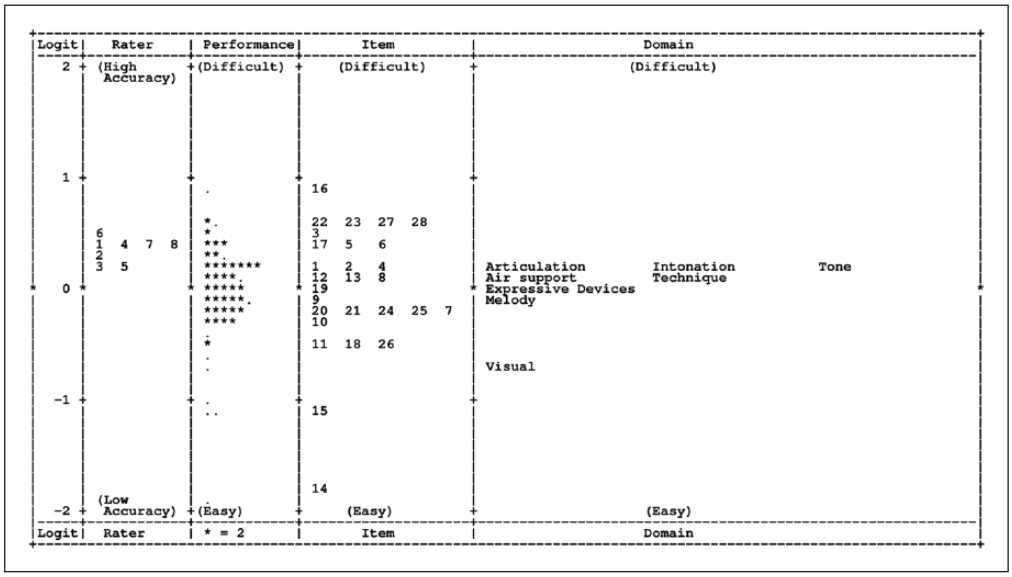

Figure 1. Variable map.

Table 3. Rater calibrations.

Rater	Average accuracy score	Measure	*SE*	Infit *MSE*	Outfit *MSE*
6	0.65	0.54	0.05	0.96	0.9
7	0.63	0.44	0.04	0.98	0.97
8	0.63	0.44	0.04	0.97	0.93
4	0.63	0.43	0.04	1.00	0.98
1	0.62	0.39	0.04	0.97	0.96
2	0.61	0.32	0.04	1.03	0.98
3	0.59	0.24	0.04	1.13	1.20
5	0.59	0.24	0.04	0.99	0.96
Mean	0.62	0.38	0.04	1.00	0.99
SD	0.02	0.11	0.00	0.06	0.09

Note. Raters are ordered by Measure, from most accurate (highest measure) to least accurate (lowest measure).

Taken together, the variable map highlights the finding that the average rater location was greater than the average location of student performances, items, and domains – suggesting that the raters who scored this music performance assessment were generally accurate. Furthermore, this plot highlights the widespread range of difficulties associated with accurately rating the individual student performances, items, and domains, and that the range of locations observed within these facets exceeds that observed within the rater facet.

Overall rater accuracy. Table 3 summarizes the calibration of the rater facet in terms of logit-scale locations and model-data fit.

Average accuracy scores are also provided for each rater in order to facilitate the interpretation of logit-scale locations. For raters, higher average accuracy scores indicate that a rater frequently provided accurate ratings, and lower accuracy scores indicate that a rater frequently

Table 4. Item calibrations.

Item	Average accuracy score	Measure	*SE*	Infit *MSE*	Outfit *MSE*
16	0.35	0.91	0.08	1.00	1.01
28	0.45	0.62	0.08	1.01	1.00
23	0.45	0.61	0.08	1.05	1.06
27	0.45	0.61	0.08	1.00	1.02
22	0.48	0.59	0.08	1.02	1.02
3	0.43	0.47	0.08	1.00	1.00
5	0.43	0.40	0.08	1.04	1.04
6	0.44	0.39	0.08	1.06	1.08
17	0.47	0.39	0.08	0.99	0.99
1	0.52	0.23	0.08	0.97	0.97
2	0.52	0.21	0.08	1.04	1.04
4	0.50	0.19	0.08	1.01	1.01
12	0.72	0.11	0.08	0.98	0.97
13	0.72	0.11	0.08	1.00	0.99
8	0.72	0.07	0.09	0.98	0.98
19	0.63	–0.05	0.08	1.03	1.05
9	0.75	–0.08	0.09	1.00	1.02
7	0.77	–0.17	0.09	0.97	0.95
20	0.66	–0.17	0.08	0.98	0.96
21	0.66	–0.20	0.08	1.02	1.01
25	0.65	–0.22	0.08	0.96	0.94
24	0.65	–0.23	0.08	1.03	1.04
10	0.79	–0.32	0.09	0.95	0.88
18	0.68	–0.50	0.08	1.01	1.00
26	0.71	–0.51	0.08	0.96	0.94
11	0.83	–0.54	0.10	0.95	0.87
15	0.89	–1.09	0.12	0.98	0.94
14	0.94	–1.83	0.17	0.99	0.86
Mean	0.62	0.00	0.09	1.00	0.99
SD	0.16	0.57	0.02	0.03	0.06

Note. Items are ordered by Measure, from most difficult to rate accurately (highest measure) to easiest to rate accurately (lowest measure).

provided inaccurate ratings. These average accuracy ratings are congruent with rater locations on the logit scale, where higher locations reflect higher accuracy. As was observed in the variable map, the logit-scale locations for the eight operational raters ranged from 0.54 logits (SE = 0.05) for Rater 6 (average accuracy score = 0.65), who was the most accurate rater to 0.24 logits (SE = 0.04) for Rater 3 and Rater 5, who were the least accurate raters (average accuracy score = 0.59). Acceptable model-data fit statistics for each of the raters suggest that these measures can be interpreted as indicators of their locations on the construct of rater accuracy.

Rater accuracy across items. Table 4 summarizes the calibration of the item facet using the same format as Table 3.

When interpreting item locations on the logit scale, it is important to note that the orientation of this facet is opposite that of the rater facet. Specifically, high average accuracy scores indicate that an item was easy to rate accurately; these easy items have lower locations on the

Table 5. Domain calibrations.

Domain	Average accuracy score	Measure	*SE*	Infit *MSE*	Outfit *MSE*
Articulation	0.43	0.25	0.08	1.04	1.04
Intonation	0.44	0.24	0.08	1.06	1.08
Tone	0.47	0.18	0.05	1.01	1.00
Air support	0.50	0.11	0.05	1.00	1.00
Technique	0.52	0.07	0.05	1.01	1.00
Expressive devices	0.56	–0.03	0.03	1.00	1.00
Melody	0.61	–0.13	0.04	1.02	1.01
Visual	0.79	–0.68	0.03	0.98	0.94
Mean	0.54	0.00	0.05	1.02	1.01
SD	0.12	0.30	0.02	0.03	0.04

Note. Domains are ordered by Measure, from most difficult to rate accurately (highest measure) to easiest to rate accurately (lowest measure).

variable map. On the other hand, low average accuracy scores indicate that an item was difficult to rate accurately; these difficult items have higher locations on the variable map. As was observed in the variable map, there is a wide range of logit-scale locations across the 28 items. Specifically, the difficulty associated with providing an accurate rating on the items ranges from 0.91 logits (*SE* = 0.08) for Item 16, which was the most difficult item to rate accurately (average accuracy score = 0.35) to -1.83 logits (*SE* = 0.17) for Item 14, which was the easiest item to rate accurately (average accuracy score = 0.94). The wide range in item difficulty, along with the high reliability of separation statistic for items, suggests that the raters who scored the MPR-2L-INSTSOLO assessment were considerably more accurate when rating student performances on some items compared to others. Additional research is needed in order to more fully understand these differences, such that rater training procedures and scoring materials can be revised to improve rater accuracy across items. Acceptable model-data fit statistics across the items suggest that these measures can be interpreted as indicators of their locations on the construct of rater accuracy.

Rater accuracy across domains. Table 5 summarizes the calibration of the domain facet using the same format as Table 3 and Table 4.

The domain facet is oriented in the same direction as the item facet. Specifically, high average accuracy scores indicate that a domain was easy to rate accurately; these easy domains have lower locations on the variable map. On the other hand, low average accuracy scores indicate that a domain was difficult to rate accurately; these difficult items have higher locations on the variable map. The difficulty associated with providing an accurate rating on the items ranges from 0.25 logits (*SE* = 0.08) for Domain 3 (Articulation), which was the most difficult domain to rate accurately (average accuracy score = 0.43) to -0.68 logits (*SE* = 0.17) for Domain 5 (Visual), which was the easiest domain to rate accurately (average accuracy score = 0.79). Similar to the item facet, the relatively wide spread of domain difficulty, along with the high reliability of separation statistic for domains, suggests important differences in rater accuracy across domains. Additional research is needed in order to more fully understand these differences in rater accuracy across domains, such that rater training procedures and scoring materials can be revised to improve rater accuracy across domains. Acceptable model-data fit statistics across the items suggest that these measures can be interpreted as indicators of their locations on the construct of rater accuracy.

Conclusions

The first research question referred to how accurate the rater judgments were, overall. Results indicated a range from 0.54 logits (*SE* = 0.05) for Rater 6 (average accuracy score = 0.65), who was the most accurate rater, to 0.24 logits (*SE* = 0.04) for Rater 3 and Rater 5, who were the least accurate raters (average accuracy score = 0.59). The second research question addressed how accurate the rater judgments were across each of the items of the rubric. Results indicated that the difficulty associated with providing an accurate rating on the items ranges from 0.91 logits (*SE* = 0.08) for Item 16, which was the most difficult item to rate accurately (average accuracy score = 0.35), to -1.83 logits (*SE* = 0.17) for Item 14, which was the easiest item to rate accurately (average accuracy score = 0.94). The third research question referred to how accurate the rater judgments were across each of the domains of the rubric. Results indicated that the difficulty associated with providing an accurate rating on the items ranged from 0.25 logits (*SE* = 0.08) for Domain 3 (Articulation), which was the most difficult domain to rate accurately (average accuracy score = 0.43), to -0.68 logits (*SE* = 0.17) for Domain 5 (Visual), which was the easiest domain to rate accurately (average accuracy score = 0.79).

Discussion and implications

The ability of test developers to produce psychometrically sound data (i.e., valid, reliable, and fair; AERA, APA, & NCME, 2014) is vitally important for any assessment context. However, in specific assessment contexts where raters mediate the assessment process, such as music performance assessment, test developers have an additional layer of subjectivity to overcome – human judgment and decision-making. Therefore, careful attention to and management of rater behavior is one of the most critical roles in providing a psychometrically sound and well-informed direct music assessment. Although divergence of response is sometimes welcomed in music performance evaluations (such as assessment contexts where the predominant purpose of the assessment is formative in nature), the results of formal music performance evaluations such as all-state auditions, formal performance evaluations, district, state, or national chair selections, college auditions, concerto competitions, or a myriad of other accountability-driven assessments are expected to be objective. Therefore, it is particularly important and relevant to implement transparent strategies to increase objectivity. In particular, methods of improving rater accuracy are especially beneficial.

As discussed in the introduction to this article, even the most qualified subject matter experts in the field of music education contribute construct-irrelevant variability to the music performance assessment context, even though in many instances, the decisions based upon music performance assessments can be high-stakes for the students being evaluated as well as the teachers associated with those students. The appropriateness of the decisions should then be based on accurate decision-making. Although there are some delimitations to this study, including the use of a small sample of student performances nested within similar classrooms, the use of only one assessment context, and the use of only video performances, the methodology in this study provides direct and verifiable evidence of the accuracy of ratings, specifically through the manner in which each rater applied the measurement instrument.

When interpreting the results from this study, it is important to note a possible limitation related to the use of a single expert rater, rather than a panel of expert raters, as a criterion against which to evaluate operational raters in terms of accuracy. Even when expert raters are directly involved in the development of scoring materials, these raters can still be subject to errors and systematic biases that may affect the interpretation of accuracy scores when they are used as criteria for defining rater accuracy. Panels of expert raters may mediate these effects,

but are still potentially subject to errors and systematic biases. In the current study, idiosyncrasies in the single expert's judgments may have influenced the difficulty ordering of the music assessment items.

Best practice in the field of music education in the United States, at large, attempts to provide safeguards in employing operational raters at various state and district performance evaluations. An evaluation of several prominent state music education association adjudicator requirements and adjudication procedures (Texas Music Adjudicators Association (n.d.), Florida Bandmasters Association (2015), Ohio Music Education Association (2008), and the New York State School Music Association (n.d.), for example), indicates several similar safeguards in selecting adjudicators. These include but are not limited to specified years of teaching experience, initial training, documented success in teaching, adjudicator experience, music degree requirements, and continued training. The problem, however, is that in all instances, focus is on behavior-centered approaches to rater behavior and not empirical-behavior centered approaches to rater behavior. Unfortunately, as evidenced in this study and other empirical-behavior centered rater behavior approaches in music performance assessment (see Wesolowski et al., 2016b for discussion of, categories of, and specific citations for rater-behavior approaches), behavior-centered approaches alone do not provide adequate evidence of rater effects to provide strong validity, reliability, and fairness argument in assessment contexts.

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 of music performance assessments. In the current study, the results indicated a wide range of difficulty in assigning accurate ratings associated with the items assessment, along with some differences in the difficulty associated with providing accurate ratings across domains. In particular, the wide range of difficulty measures for items and domains, along with the high values of reliability of separation statistics, suggest that some items and domains are substantially easier to rate accurately than others. In order to support the interpretation and use of ratings from the MPR-2L-INSTSOLO assessment, it is necessary to investigate these differences in rater accuracy in order to more fully understand the specific aspects of the assessment that are contributing to raters' difficulty in providing accurate ratings, such as the structure of the item, performance level descriptors, or insufficient training. Rater training procedures can then be adjusted to more fully address these difficult-to-rate aspects of the assessment.

Considering these results more broadly, what is currently missing from controlling rater quality in music performance assessment are thorough rater training and recalibration processes. Wolfe et al. (2015) describe the processes of rater training and the role of accuracy indices in the context of writing assessment:

> ... raters are engaged in a training process, which may last several days, in which they review the purpose of the assessment and scoring process, thoroughly review the scoring rubric, review annotated and scored examples of examinee responses, and practice applying the scoring rubric to additional training examples. At the completion of training, raters often must demonstrate a mastery of the scoring rubric by achieving a pre-determined level of agreement with true scores (i.e, expert consensus scores) that have been assigned to responses that are collected into a qualifying test. Raters who do not achieve the required minimum level of mastery either receive additional training or are released from the scoring project (p. 154).

As indicated by Wolfe, the use of rater accuracy indices in rater training processes is used to certify raters' mastery of the rubric use and the specified standards for the operational scoring process. He further continues to describe the rater recalibration process:

> Beyond this initial training period, raters are also recalibrated periodically (i.e, brief retraining activities) in order to ensure that their scores remain consistent with the intentions of the scoring rubric. Additionally, several operational rater monitoring procedures are employed to confirm that the scores that are reported to examinees do not deviate from that of the standard of rating quality (p. 154)

In revisiting Hash's (2013) statement, "It is unclear why ratings trend toward the top of the scale" (p. 165), one can speculate as to several process-based reasons why scoring is negatively skewed: (a) lack of validity evidence of the functioning of the measurement instrument itself, (b) use of scoring procedures based upon Classical Test Theory that confuse raw scores for linear measures, (c) lack of empirical evidence of rater behavior, (d) lack of standards-based scoring procedures, (e) lack of rater training where clarity in the use of the instrument and knowledge of performance standards are clear, and (e) lack of rater recalibration throughout continued assessment contexts. Until the field of music education accepts the impact of its psychometric and rater negligence, music performance assessments will continue to be limited in terms of their validity, reliability, and fairness.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

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

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

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

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

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

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

Florida Bandmasters Association. (2015). Florida Bandmaster Association adjudication manual. Retrieved from http://fba.flmusiced.org/media/1274/adjudication-manual-2015.pdf

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

Hash, P. M. (2013). Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review*, *114*(4), 163–169.

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.

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

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.

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

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619–624.

Murphy, K., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective*. Boston, MA: Allyn & Bacon.

National Association for Music Education (NAfME). (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/

New York State School Music Association. (n.d.). Becoming a NYSSMA adjudicator. Retrieved from http://www.nyssma.org/committees/adjudication-festival-committee/director-of-adjudicators/

Ohio Music Education Association. (2008). *Rules and regulations for OMEA adjudicated events*. Retrieved from http://www.omea-ohio.org/Static_PDF/REMOVED%20PDFs/AE/OMEA_Rulebook.pdf

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Texas Music Adjudicators Association (n.d.). Membership requirements. Retrieved from http://www.txmaa.org/tmaameminfo.php

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

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. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, ..., & Williams, J. D. (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. (revision under review). Judgmental standards setting: The development of objective content and performance standards for secondary-level solo instrumental music assessment.

Wesolowski, B. C., Burrack, F., & Parkes, K. A. (in press). Phenomenography: Bringing together theory and practice through the process of national standards development and measure construction. In T. S. Brophy & M. Fautley (Eds.). *Context matters: Selected papers from the 6th International Symposium on Assessment in Music Education*.

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

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

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

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

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299.

Wind, S. A., Engelhard, G., & Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment*, *21*(4), 278–299.

Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, *16*(2), 153–160.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1–10.

Appendix 1. 28-item Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO; Wesolowski, 2017).

Technique				
1. *Finger/slide dexterity*	• Agility poses an extreme barrier to efficiency in fingers/slide when changing notes. • Fingers consistently exhibit tension in lyrical and/or technical passages. • Note changes are a serious problem in performance.	• Agility poses a moderate barrier to efficiency in fingers/slide when changing notes. • Fingers often exhibit some tension in lyrical and/or technical passages. • Note changes are a moderate problem in performance.	• Agility poses no barrier to efficiency in the fingers/slide when changing notes. • Fingers rarely exhibit tension during lyrical or technical passages. • Note changes are not a problem in performance.	
2. *Coordination between tongue and fingers/slide*	• Tongue and fingers/slide are rarely coordinated during performance. • Gaps between finger/slide changes and articulation initiation consistently detract from the performance.	• Tongue and fingers/slide are sometimes uncoordinated. • Gaps between finger/slide changes and articulation initiation consistently often detract from the performance.	• Tongue and fingers/slide are consistently coordinated. • Gaps between finger/slide changes and articulation initiation consistently rarely detract from the performance.	
Tone				
3. *Tone quality in varying registers*	• The quality of sound is rarely characteristic when exchanges occur between registers. • Changes in tone quality between registers are a serious problem during performance.	• The quality of sound is sometimes characteristic when exchanges occur between registers. • Changes in tone quality are a moderate problem during performance.	• The quality of sound is often characteristic when exchanges occur between registers. • Changes in tone quality are a minor problem during performance.	• The quality of sound is consistently characteristic when exchanges occur between registers. • Changes in tone quality are not a problem during performance.

(Continued)

Appendix 1. (Continued)

4. *Tone while executing expressive gestures*	• Command of characteristic tone is consistently compromised while executing expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.	• Command of characteristic tone is sometimes compromised through various combinations of expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.		• Command of characteristic tone is rarely compromised through various combinations of expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.
Articulation				
5. *Consistency of articulation*	• Articulations are often inconsistent in passages with notes of a similar style and detract much from the performance. • It is typical for notes to be unnecessarily accented or stressed by inconsistent attacks.	• Articulations are sometimes inconsistent in passages with notes of a similar style but detract very little from the performance. • A few notes are unnecessarily accented or stressed by inconsistent attacks.		• Articulations are rarely inconsistent in passages with notes of a similar style and do not detract from the performance. • Notes are not accented or stressed by inconsistent attacks.
Intonation				
6. *Intonation accuracy*	• Intonation accuracy is a serious problem.	• Intonation accuracy is a moderate problem.	• Intonation accuracy is a minor problem.	• Intonation accuracy is not a problem.
Visual				
7. *Body posture*	• Upper body carriage alignment (i.e., shoulder, spine) inappropriate for instrument.		• Upper body carriage alignment (i.e., shoulder, spine) appropriate for instrument.	
8. *Instrument angle*	• Instrument angle is inappropriate for instrument.		• Instrument angle is appropriate for instrument.	
9. *Head position*	• Head alignment is inappropriate for instrument.		• Head alignment is appropriate for instrument.	
10. *Arm position*	• Arm position is inappropriate for instrument.		• Arm position is appropriate for instrument.	
11. *Wrist position*	• Wrist position is inappropriate for instrument.		• Wrist position is appropriate for instrument.	

Appendix 1. (Continued)

12. Hand position	• Hand position is inappropriate for instrument.		• Hand position is appropriate for instrument.
13. Embouchure/ flexibility	• Embouchure is inappropriate for instrument. • Embouchure detracts from musical performance.		• Embouchure is appropriate for instrument. • Embouchure does not detract from musical performance.
14. Cheeks	• Cheeks are puffy and detract from embouchure support and airflow.		• Cheeks are not puffy and do not detract from embouchure support and air flow.
15. Jaw movement	• Extraneous jaw motion is consistently present in articulation.	• Extraneous jaw motion is sometimes present in articulation.	• Extraneous jaw motion is rarely present or not present in articulation.
Air Support			
16. Breath intake	• Breath intake is rarely full, deep, or initiated from the diaphragm (i.e. breathing through the nose, breathing through the instrument, shallow breathing initiated from the chest/shoulders).	• Breath intake is often full, deep, and initiated from the diaphragm (i.e. some breathing through the nose, some breathing through the instrument, some shallow breathing initiated from the chest/ shoulders).	• Breath intake is consistently full, deep, and initiated from the diaphragm.
17. Sufficiency of air	• Air support often inconsistent and detracts much from the quality of the performance.	• Air support is sometimes inconsistent and detracts very little from the performance.	• Air support is rarely inconsistent and does not detract from the performance.
18. Air support in various registers of the instrument	• Tone is inappropriately supported at various registers of the instrument.		• Tone is appropriately supported at various registers of the instrument

(Continued)

Appendix 1. (Continued)

Melody				
19. *Note accuracy*	• Student often demonstrates note inaccuracy.	• Student sometimes demonstrates note inaccuracy.	• Student rarely demonstrates note inaccuracy.	
20. *Communication of musical phrases*	• Phrases are inappropriately contoured.		• Phrases are appropriately contoured.	
21. *Connection of phrases*	• Does not meaningfully connect phrases.		• Meaningfully connects phrases.	
22. *Inflection at cadence points*	• Melodic line rarely demonstrates inflection at cadence points.	• Melodic line sometimes demonstrates inflection at cadence points.	• Melodic line consistently demonstrates inflection at cadence points.	
Expressive Devices				
23. *Stylistically-related dynamics*	• Dynamics are rarely appropriate for the style of music being performed.	• Dynamics are sometimes appropriate for style of music being performed.	• Dynamics are often appropriate for the style of music being performed.	• Dynamics are consistently appropriate for the style of music being performed.
24. *Contrast in dynamics*	• Rarely demonstrates meaningful contrast in dynamics.	• Student sometimes demonstrates meaningful contrast in dynamics.	• Student frequently demonstrates meaningful contrast in dynamics.	
25. *Subdivision of the rhythm*	• Inaccurate performance of subdivisions detracts from solidly communicated tempo and meter.		• Accurate performance of subdivisions contributes to solidly communicated tempo and meter.	
26. *Appropriateness of tempo*	• Tempo is inappropriate throughout the performance.		• Tempo is appropriate throughout the performance.	
27. *Steadiness of pulse*	• Control of pulse detracts very much from the continuous flow of the music.	• Control of pulse sometimes detracts from the continuous flow of the music.	• Control of pulse does not detract from the continuous flow of the music.	
28. *Expressive pulse and tempo fluctuations*	• Expressive changes in tempo and pulse are inappropriate for the style and demands of the literature.	• Expressive changes in tempo and pulse are slightly inappropriate for the style and demands of the literature.	• Changes in tempo and pulse are appropriate for the style and demands of the literature.	