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

Validation of a String Performance Rubric Using the Multifaceted Rasch Measurement Model

Kinsey E. Edwards
The University of Georgia
Athens, GA

Andrew S. Edwards
The University of Georgia
Athens, GA

Brian C. Wesolowski
The University of Georgia
Athens, GA

ABSTRACT

*The purpose of this study was to develop a valid and reliable rubric for the evaluation of large ensemble string performances using psychometric principles of invariant measurement, whereby the measurement of ensembles, items, and raters are simultaneously and independently calibrated. The multifaceted Rasch partial credit measurement model was used in order to achieve invariant measurement. This study was guided by the following research questions: (a) What does Rasch measurement analysis reveal about the psychometric quality (i.e., validity and reliability) of items, raters, and ensembles within the context of a large ensemble string performance assessment? (b) How do the items vary in difficulty, raters vary in severity, and ensembles vary in achievement? (c) How does the rating scale structure vary across individual items? Music content experts (*N = *25) were solicited to each evaluate 4 string ensemble performances. A 4-point Likert-type scale (e.g., strongly agree, agree, disagree, and strongly disagree) was used to rate performances. An incomplete rater assessment network was used wherein a total of 52 performances were evaluated. Results indicated that 27 items demonstrated fit to the measurement model. Response categories for each of the items were optimized resulting in a range of 2 to 4 performance criteria in order to increase measurement accuracy and precision. Implications for the improvement of music assessment practices are discussed.*

As the educational setting becomes more data driven, valid, reliable, and fair empirical evidence is needed to demonstrate growth in student achievement (Brookhart, 2013). Additionally, the trend of measuring student achievement using valid and reliable empirical data continues to become more prominent with continued focus on teacher effectiveness (Brookhart, 2013). In the field of music, however, the resulting data that

is used for these purposes is often psychometrically misleading due to the current misalignment between instructional focus and corresponding assessment methods (Colwell, 2003).

The results from evaluations in academic classroom settings (such as math and science) are typically student interaction measures that can be enumerated easily through the use of cognitive tests including multiple-choice, true/false, and other selected-response-type examinations (Blakeslee, 2004). In contrast, assessment results in music performance settings are often given in the form of a qualitative narrative that addresses how well parts of a whole work together to contribute to the final performance (Hope & Wait, 2013). Assessment measures in a music classroom are best achieved through the use of rater-mediated evaluations that allow for such narrative and critique to be shared (Wesolowski, 2012). Most often, in order to cater to the data-driven focus in nonperformance-based classrooms, the approach implemented by music teachers is the use of selected-response testing similar to academic classrooms (Hope & Wait, 2013). These methods allow for only one clear answer, thereby allowing for empirical data to be provided more easily. Assessment practices should instead be considerate of authentic behaviors in the context of music teaching and learning, where students demonstrate performance-based tasks that are relevant to the content area and the manner in which the content is being delivered (Zaleski, 2014).

Regardless, music educators must present empirical evidence to document levels of student achievement despite the performance-based nature of music (Hope & Wait, 2013). In the music classroom, the instructional focus requires spending time to develop performance skills while also developing the ability to decipher how to execute and make artistic decisions based on those performance skills (Hope & Wait, 2013). Evidence of student learning therefore needs to be gathered by assessing the performance knowledge and skills that are being taught (Brewer, Knoeppel, & Lindle, 2014).

Examples of current music performance evaluations that are used to evaluate student performance achievement include juries, auditions, chair placements, large group performance competitions, and community and public performances (Hope & Wait, 2013). One particular challenge arises in the use of student performance opportunities for empirical evidence of achievement due to the fact that such observations of student performance are rater mediated (Wesolowski, Wind, & Engelhard, 2015). The rater interaction in such situations allows for multiple perspectives and opinions that may affect the consistency of feedback (Hope & Wait, 2013). Unlike selected response standardized assessments used in more academic-type classrooms, validity issues in music assessment practices occur because these measures are not indicative of direct student interaction, but rater interaction instead. Thus, even if the performer and performance remains constant, a different evaluator may very easily yield entirely different results each time. The nature of the performance-based, rater-mediated assessment unfortunately allows for subjectivity to occur on the part of the rater (Wesolowski, Wind, & Engelhard, 2016a). For this reason, the field of music education must continue working

to develop valid, reliable, fair, and, more importantly, authentic performance evaluation tools in order to replace the current measures that are being used (McMillan, 2003). The purpose of this study is to help meet this immediate need by developing a performance measure that is authentic to the teaching and learning in the classroom and that has been tested for validity and reliability the same way that high-stakes student interaction measures (such as open response writing assessments) are validated. In order to make valid and reliable inferences of student achievement in a performance-based music classroom, it is important to empirically investigate and document the measure construction process.

THE USE OF RUBRICS IN MUSIC TEACHING AND LEARNING

One solution to the misalignment between teaching and assessment practices in the music profession can be achieved through the use of valid and reliable rubrics, as rubrics can accurately account for multiple technical and expressive aspects of music performance (DeLuca & Bolden, 2014). An understanding of what a rubric is and why it is an effective measure of music performance is required in order to entertain a discussion on assessment practices in the music discipline. According to Asmus (1999), rubrics can be defined as "a set of scoring criteria used to determine value of a student's performance on assigned tasks; the criteria are written so students are able to learn what must be done to improve their performance in the future" (p. 21). The information and criteria presented in a rubric helps to ensure that both the student and teacher are informed as to the direction and expectations that will materialize in the classroom, thereby providing a better-established communication link between the teacher and student (Whitcomb, 1999). Rubrics also provide a method for documentation of student achievement levels wherein specific written feedback can be shared with teachers, parents, and students in order to provide evidence of student performance achievement in the music classroom (Wesolowski, 2012). The effective use of rubrics in music performance settings can help students to develop ownership and to take control of their learning efforts (DeLuca & Bolden, 2014).

Research efforts in the music education community have contributed to an increased understanding of how rubrics can improve scoring reliability and consistent grading methods for the music classroom (DeLuca & Bolden, 2014). However, there is still an urgent need for the development of valid and reliable rubrics that can be used to consistently yield results that adequately measure student performance in large group performance evaluation settings and that can also help to inform preparations for such in the music classroom (Colwell, 2003).

The misalignment between musical performance abilities and resulting assessment data is also part of a larger problem that stems from a lack of teacher training focused on understanding true assessment methods (Colwell, 2003). Parkes (2010) states, "Few

educators received any formal training in assigning marks to students' work or in grading students' performance and achievement" (p. 98). As a result, many music educators do not fully understand the true functions and possibilities that can result from using valid and reliable rubrics in the music classroom (Wesolowski, 2012). The unfortunate reality that further confounds this issue is that teachers are often given the task of developing rubrics with very little training (Pellegrino Conway, & Russell, 2015). This greatly limits the ability of music educators to accurately measure student achievement in a valid and reliable way that can be used to help inform teaching and learning practices (DeLuca & Bolden, 2014).

Due to the current focus on assessment and accountability at the national level, the notion of developing a system of more authentic measures for the music classroom is gaining thought and attention (Model Cornerstone Assessments, 2015). A change in the way in which the field of music assesses musical behaviors in favor of a more validated approach will allow for music educators to accurately and confidently report evidence of student learning to students, parents, colleagues, and administrators (Pellegrino et al., 2015). Teachers should confirm that the rubrics are valid in that they measure what is intended to be measured and also that they are reliable in providing consistent results throughout the assessment process (Pellegrino et al., 2015).

Future implications of this measurement system are specific to the validation of a new system that can be used by classroom and university music educators to rate large group string ensemble performances. Achievement parameters and intended goals of string performance need to be constructed and agreed upon prior to the development and implementation of performance evaluations. This contributes to the idea that rubrics should only be used in the classroom for performance preparations if they are first rigorously tested. This is especially true if the data from these assessment measures is going to be used as a means to infer teacher effectiveness. Messick (1989) refers to this consideration as consequential validity. In this instance, the results that are used from performance evaluations have important implications and social consequences that must be considered prior to implementing such high level forms of assessment.

In order for the results of the rubrics to maintain meaning as a part of the instructional process, focused efforts must be directed toward the development and validation of rubrics (DeLuca & Bolden, 2014). Though there is an increased focus on the use of theoretically informed rubrics in student performance assessment, there is still a need for the development of empirically supported rubrics (DeLuca & Bolden, 2014). The challenge in using rubrics as an assessment mechanism for string ensemble performance lies in the notion that raters mediate current assessments in music as a mechanism for providing the empirical data. As such, these measures do not take into consideration various rater errors, such as severity/leniency and raters' specific use of the rating scale structure. This leads to subjectivity on the part of the rater, thereby providing data that does not accurately measure the true level of student achievement (Wesolowski, Wind, & Engelhard, 2016a, 2106b).

The proposed method therefore favors the consideration of a psychometric approach to developing such rubrics. An advantage of using such an analysis approach is that the metrics allow us to gain a better understanding for (and infer) local independence prior to the use of the developed rubric (Linacre, 2010). There is strong evidence that a measure that is put through rigorous testing outside of the classroom (i.e., in a large ensemble performance evaluation setting) will hold its validity when used in the true testing environment (i.e., the performance-based music classroom).

PSYCHOMETRIC CONSIDERATIONS

Arguably, the most significant limitation of performance assessments is measurement variance attributed to raters. Rater scores are less associated with the performances themselves and more associated with the perceptive lens of the rater (Brunswik, 1952; Engelhard, 2002). This is true in any rater-mediated situations; however, this specific scenario requires an examination of the nature in which individuals rate string ensemble performance. Traditional methods of evaluating rater behavior in music include consistency and consensus estimates. These methods do not adequately estimate true scores of performances (Wesolowski et al., 2015). More specifically, raters can consistently overestimate or underestimate true scores but demonstrate high consistency and consensus estimates. Inferences drawn from such instances can therefore be misleading. In order for rater-mediated assessment processes to be more fair, rater errors, such as severity/leniency, need to be investigated as part of the measurement process (Wesolowski et al., 2016a).

The development of a valid performance evaluation therefore requires consideration of the psychometric process in which a rating system can be constructed in a way that will allow for it to be applied to future performance scenarios. The multifaceted Rasch partial credit measurement model (Linacre, 1989) was used in this study to investigate the psychometric properties (i.e., validity and reliability) of the original rating scale because of the properties of invariant measurement underscoring the Rasch family of measurement models (Engelhard, 2013). Content validity is considered through the discarding of items that are not considered to be useful in the measurement of string ensemble performance. This notion is specifically referred to as data-model fit when using the Rasch model (Wesolowski et al., 2016a, 2016b). When using the Rasch model, instead of the model of a normal bell curve being mapped to the data, the data is instead compared to an already existing and consistent model. Any data that does not adequately fit the model based upon the properties of invariant measurement is discarded. Data-model fit will be determined by evaluating fit indices for all items.

When using the Rasch model, data-model fit is determined based on the degree to which invariant measurement is met. Adequate fit to the model results when the five requirements for invariant measurement are met (Engelhard & Perkins, 2011): (a) the items must be independent of the persons used for measurement (i.e., person-invariant

calibration of items); (b) the persons must have a higher probability of success on easy items in comparison to the more difficult items (i.e., noncrossing person response functions); (c) the persons must be independent of the items used for measurement (i.e., item-invariant calibration of persons); (d) a person who is more able must have a higher probability on succeeding on more difficult items than that of a less able person (i.e., noncrossing person responses); and (e) items must measure a single underlying latent variable (i.e., Engelhard, 2013). When adequate fit to the model is obtained, invariant measurement is achieved.

In contrast to factor analysis methods, when using Rasch, there is no conflict between the observed data and the future use of the model. Due to the independent measurement of raters, performances, and items, the model is sample independent and can therefore be applied to future assessments (Meredith, 1993). This aspect of the Rasch model accounts for reliability. The developed rating scale can be applied to future performances and will yield consistent results because invariant measurement was used.

The partial credit component of the model (Masters, 1982) allows for the additional parameter of rating scale structure to be explored across each item. In the context of music performance assessment, evidence exists that each of the categories (e.g., strongly agree, agree, disagree, strongly disagree) for each item are not equidistant because they vary in difficulty in terms of ability to endorse (Wesolowski et al., 2016b). The partial credit aspect of the Rasch model allows for the investigation of the difficulty level across each rating scale category. For instance, a strongly agree is harder to achieve than agree. The partial credit allots for this varying degree of difficulty because a higher level of achievement should be earned with strongly agree, as opposed to agree. To investigate the partial credit aspect of the model, consideration of monotonicity (e.g., proper ordering of rating scale categories) between categories, appropriate distinction made between performances, frequency of use by raters, and probability measures were taken into account in order to determine the most optimized structure for each of the rating scale categories (Linacre, 2002). If the logit measurements showed that agree was harder to earn than strongly agree, that would result in a violation of monotonicity. Analysis of the rating data was conducted using the computer program FACETS (Linacre, 2014).

In this data-driven educational climate, there is a critical need for the development of valid, reliable, and fair measures that can be used to measure student achievement in string ensemble performance settings. Such assessment measures need to also be applied in performance-based music classrooms by providing information that can help to guide instructional decisions that are implemented as a component of preparations for such large ensemble evaluations. The purpose of this study was to develop a valid and reliable rubric for the evaluation of large ensemble string performance. This study was guided by the following research questions:

1. What does Rasch measurement analysis reveal about the psychometric quality (i.e., validity and reliability) of items, raters, and ensembles within the context of a large ensemble string performance assessment?

2. How do the items vary in difficulty, raters vary in severity, and ensembles vary in achievement?
3. How does the rating scale structure vary across individual items? (The null hypothesis states that the final items on the rating scale will share identical response structure.)

Method

Rater cohort of content experts. Twenty-five content experts participated in this study by agreeing to listen to and evaluate four full ensemble orchestra recordings each. Fifteen females and 10 males participated in the study, 20 of whom attained a bachelor's, master's, or specialist's degree and five of whom attained their doctoral degree. Of the 25 content experts, 15 teach in a middle school string setting, and 10 teach in a high school string setting. These content experts will benefit from the developed rubric in that the resulting findings can be used to evaluate performances by their programs. Each rater was chosen based on their availability, experience, influence in the field, and willingness to listen to the recordings and rate the performances. The selection of the content experts was based on the assumption that "best practice in the selection and utilization of adjudicators in the field of music performance suggests that expert teachers and performers offer the best chance for providing a fair and equitable assessment" (Wesolowski et al., 2015, p. 165).

Development of initial item pool. Thirty-eight item stems were extracted from an original item pool previously developed by Zdzinski and Barnes (2002). Zdzinski and Barnes used some stems from the same item pool to develop an earlier rating scale for string performance. The treatment of the item stems from this pool was different in the original study from which they were extracted because factor analysis was used. When using the factor analysis method, individual characteristics are not independent of one another and therefore the resulting rating scales cannot be applied to future situations.

The descriptive statements were organized into four a priori categories based upon the performing dimension of the National Association for Music Education Model Cornerstone Assessment: (a) tone production, (b) rhythm and pulse accuracy, (c) pitch and intonation accuracy, and (d) expressive qualities/stylistic interpretation (National Association for Music Education, 2015). Three content experts reviewed each of the original item stems in order to evaluate the manner in which the stem was able to accurately describe the music concept. Discussions resulted in the editing and adapting of stems that were not considered to be clear and appropriate for the study. Agreement was reached in the directionality of the items, resulting in 20 positively phrased items and 18 negatively phrased items. The 38 items were randomized and paired with a four-point Likert-type scale (see Appendix A here: http://bcrme.press.illinois.edu/media/215/).

Performance stimuli. The content experts evaluated a total 52 recordings from a formal district music performance assessment in a large southern state that occurred in

the previous year. These recordings included string ensemble performances from both middle school and high school groups of various ability levels. These performances were representative of the population that will benefit from the development of the rubric. All recordings were professionally created and matched in sound quality.

Rater assessment network. An incomplete assessment network was used where a total of 52 performances were evaluated. Each rater listened to and evaluated a total of four performances, but two of those performances overlapped with the subsequent rater. The last rater and first rater were overlapped in order to account for all performances (Engelhard, 1997). Performances were randomly assigned to raters and were shared with individual Dropbox links. Raters used a separate randomized Google form in order to submit evaluations of each performance. Item stems were randomly presented on each Google form in order to control for rater fatigue. Once completed, negatively phrased items were reverse coded prior to analysis (see note at the bottom of Table 3 for stems that were reverse coded).

Results

Variable map. The variable map is a visual representation of the latent construct (e.g., large ensemble string performance). Each of the facets included in the study are displayed in each of the columns on the variable map. The first column shows the logit scale that serves as a "ruler" in order to allow for the measurement of each facet to be shown on a common map. The second column shows the performances, notated through the use of an asterisk for each performance. The performances near the top are considered to be the highest achieving performances and those closer to the bottom are the lower achieving performances. The measures ranged from –1.81 logits to 2.56 logits with a demonstrated range of 4.37 logits (M = –0.02, SD = 1.01, N = 50). The third column represents the severity of raters. Severity and leniency ranged from –2.19 logits to 2.07 logits with a demonstrated range of 4.26 logits (M = 0.00, SD = 0.88, N = 25). The raters closest to the bottom of the map are considered to be more lenient and raters closer to the top are considered to be more severe in their measurement practices. The fourth column shows the difficulty to endorse each item. Difficulty ranged from –1.18 logits to 1.68 logits with a demonstrated range of 2.86 logits (M = 0.00, SD = 0.62, N = 38). The items closer to the bottom of the map are considered to be easier to endorse and items closer to the top are considered to be more difficult to endorse. The measurement of these three facets will be used to infer measurement on the latent construct of large ensemble string performance.

The variable map provides a visual representation of the information that is needed to answer the second research question. Psychometric aspects of the model allow for the investigation as to how well the items, raters, and ensembles fit the model. In this particular investigation, any items that did not fit the model were discarded. Items that were considered too easy or too difficult to endorse will not be included in the final validated scale and will therefore also not be included as a part of the final rubric. The

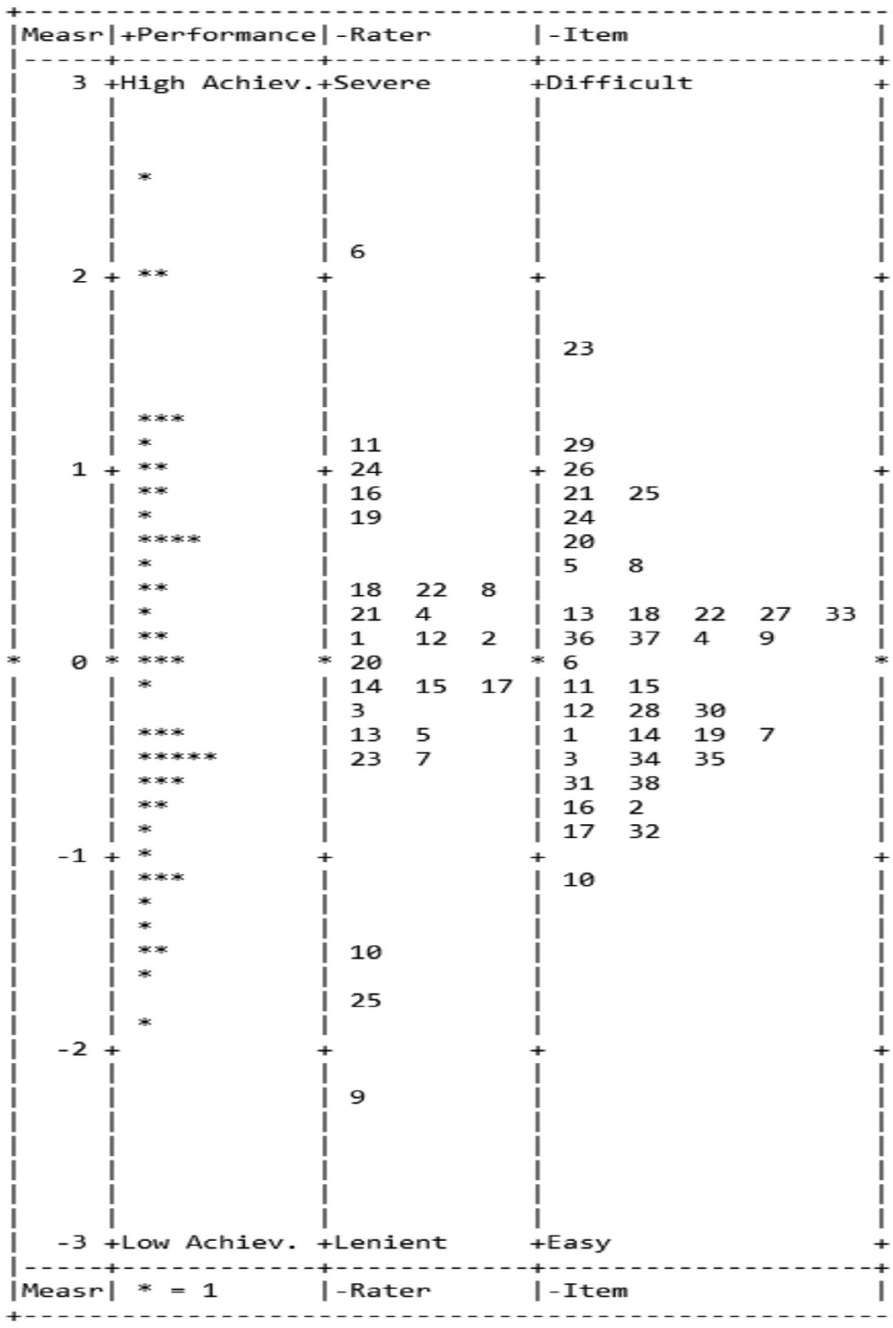

Figure 1. Variable map.

use of the Rasch model allows for any unfit items to be removed in order to aid in the creation of a valid and reliable final rating scale. The final rating scale was then translated into a rubric that will be useable to professionals in the music education setting hoping to rate string ensemble performances to show levels of student achievement.

The following calibration details explain the intricacies of how items were either considered to fit the model or considered as being misfit. Infit *MSE* statistics considered to fit the model are within the range of 0.80 and 1.20 logits as indicated by Wright and Linacre (1994) and Engelhard (2009). Measurements below 0.80 are considered to be underfit, and any measurements above 1.20 are considered to be overfit. Underfit and overfit items are considered to be misfit when applied to the model.

Calibration of ensemble performances. The calibration of student performances is provided in Appendix B here: http://bcrme.press.illinois.edu/media/215/. Higher numbers represent higher performance achievement and lower measures represent lower performance achievement. Performance 5 represented the highest performance achievement (2.56 logits) and Performance 3 represented the lowest performance achievement (–1.81 logits). Misfitting performances are based upon infit *MSE* statistics that fall outside of the ranges of 0.80 and 1.20 logits as indicated by Wright and Linacre (1994) and Engelhard (2009). Overfitting performances include Performances 2, 3, 10, 15, 17, 18, 21, 26, 31, 41, and 47. Underfitting performances include Performances 4, 6, 9, 11, 14, 19, 23, 24, 25, 27, 28, 30, 32, 36, 39, and 45.

Calibration of raters. The calibration of raters is provided in Appendix C here: http://bcrme.press.illinois.edu/media/215/. The table demonstrates a ranking of the raters in terms of severity and leniency. Rater 6 was the most severe (observed average = 1.77, logit measure = 2.07) and Rater 9 was the least severe (observed average = 2.96, logit measure = –2.19). Raters 2, 4, 7, 8, 10, 12, 13, 14, 15, 20, 21, and 22 were considered to demonstrate muted patterns with infit *MSE* less that 0.80. Raters 1, 5, 9, 16, and 25 were considered to demonstrate sporadic patterns with infit *MSE* greater than 1.20. This aspect of the model accounts for rater behaviors, which will provide pertinent information for future rater training.

Calibration of items. The calibration of items is presented in Table 1. The calibration of items displays the difficulty of each item. The more difficult items are evident in the larger logit measures, and the easier items are evident in the smaller logit measures. The most difficult item was Item 23 ("ensemble performs with consistently good intonation in all registers"; observed average = 1.95, logit measure = 1.68) and the easiest item was Item 10 ("tempi are appropriate for style of composition"; observed average = 3.14, logit measure = –1.18). Items that demonstrated overfit included Items 2, 10, 15, and 29. Items that demonstrated underfit included Items 1, 3, 4, 5, 22, 33, and 34. This provided grounds for removal from the final rating scale, which also meant that these items would not be used in the development of the rubric. Misfit items do not adequately contribute to the rating of string performance evaluation, so they should not be kept in

Table 1
Calibration of the Item Facet

Item number	Observed average	Measure	SE	Infit MSE	Std. infit	Outfit MSE	Std. outfit
23	1.95	1.68	0.16	0.84	–1.30	0.89	–0.70
29	1.91	1.18	0.16	1.21	1.40	1.29	1.80
26	2.06	0.98	0.15	0.98	–0.10	1.00	0.00
25	2.11	0.85	0.15	1.16	1.20	1.23	0.60
21	2.14	0.84	0.15	0.80	–1.60	0.84	–1.10
24	2.16	0.81	0.16	0.97	–0.10	1.00	0.00
20	2.28	0.61	0.16	0.87	–1.00	0.87	–1.00
5	2.28	0.54	0.17	0.75	–1.90	0.75	–1.90
8	2.41	0.53	0.17	0.99	0.00	0.98	0.00
33	2.52	0.27	0.10	0.74	–1.90	0.71	–2.10
27	2.36	0.23	0.13	1.07	0.50	1.11	0.80
18	2.44	0.23	0.17	1.10	0.70	1.11	0.70
13	2.46	0.21	0.16	0.91	–0.60	0.96	–0.20
22	2.48	0.20	0.16	0.67	–2.70	0.67	–2.60
4	2.48	0.17	0.16	0.75	–1.90	0.74	–2.00
9	2.49	0.09	0.17	0.94	–0.40	0.94	–0.40
37	2.50	0.09	0.15	1.14	1.00	1.18	1.20
36	2.45	0.06	0.15	0.86	–1.00	0.88	–0.80
6	2.52	0.06	0.15	0.93	–0.50	0.94	–0.40
15	2.58	–0.08	0.18	1.34	2.20	1.39	2.40
11	2.60	–0.18	0.17	1.14	1.00	1.20	1.30
28	2.67	–0.25	0.15	0.97	–0.10	0.94	–0.30
30	2.62	–0.27	0.16	1.11	0.80	1.13	0.90
12	2.60	–0.30	0.17	0.94	–0.30	0.94	–0.30
7	2.65	–0.35	0.17	0.93	–0.50	0.93	–0.40
19	2.71	–0.35	0.17	1.20	1.30	1.23	1.50
1	2.68	–0.36	0.16	0.71	–2.30	0.71	–2.30
14	2.67	–0.38	0.16	1.17	1.20	1.21	1.40
35	2.56	–0.53	0.10	0.92	–0.50	0.94	–0.40
3	2.70	–0.53	0.18	0.77	–1.70	0.73	–1.90
34	2.75	–0.55	0.18	0.77	–1.60	0.74	–1.80
31	2.76	–0.57	0.17	1.07	0.40	1.12	0.80
38	2.77	–0.61	0.17	1.15	1.00	1.12	0.80
16	2.89	–0.72	0.18	1.06	0.40	1.06	0.40
2	2.82	–0.73	0.17	1.38	2.30	1.37	2.20
17	2.87	–0.83	0.17	0.96	–0.20	0.90	–0.60
32	2.85	–0.83	0.17	0.93	–0.40	0.91	–0.60
10	3.14	–1.18	0.17	1.47	2.20	1.54	2.60
Mean	2.52	0.00	0.16	0.99	–0.10	1.00	0.00
SD	0.27	0.62	0.01	0.19	1.30	0.21	1.40

Note: The items are presented in measure order from most difficult to least difficult.

the rubric as a means of defining exemplary string ensemble performance evaluation. Further analysis could provide the opportunity to investigate additional stems that would replace the gaps represented by the removal of these stems.

Summary statistics. Summary statistics are provided in Table 2. Analysis indicates significant differences between performances (χ^2 =1383.0, $p < .01$), raters (χ^2 =1030.1, $p < .01$), and item stems (χ^2 = 542.6, $p < .01$). Good data fit is evident in that the mean square fit values (infit *MSE* and outfit *MSE*) are close to the expected value of 1.00. Acceptable range for productive parameter-level mean square statistics is between 0.80 and 1.20, according to Wright and Linacre (1994) and Engelhard (2009). Therefore, the reliability of separation for performances ($Rel_{performances}$ = .97), raters (Rel_{raters} = .98), and items (Rel_{items} = .93), shows an adequate amount of separation to confirm the construct validity of the measurement instrument. This might be more clearly understood by saying that there is 97% reliability that this assessment tool distinguishes the level of achievement of each of these performances. Thus, the final rating scale can be considered valid because the results are independent from the performances, raters, and items used to construct the rating scale. Table 2 provides information that can be used to provide the analysis necessary to answer the first research question.

Table 2
Summary Statistics from the PC-MFR Model

	Facets		
	Performance	Rater	Item
Measure (logits)			
Mean	–0.02	0.00	0.00
SD	1.01	0.88	0.62
N	50	25	38
Infit MSE			
Mean	0.99	1.00	0.99
SD	0.38	0.44	0.19
Std. infit MSE			
Mean	–0.30	–0.40	–0.10
SD	2.30	3.50	1.30
Outfit MSE			
Mean	1.00	1.00	1.00
SD	0.39	0.43	0.21
Std. outfit MSE			
Mean	–0.20	–0.30	0.0
SD	2.30	3.50	1.40
Separation statistics			
Reliability of separation	0.97	0.98	0.93
Chi-square	1383.0*	1030.1*	542.6*
Degrees of freedom	49	24	37

*$p < 0.01$

Rating scale category diagnostics. The original 38 item stems were extracted from an original item pool previously developed by Zdzinski and Barnes (2002). Following the study, misfit items were removed from the item pool and the rating scale was closely studied in order to determine the best structure for the remaining items (Linacre, 2002). In order to improve validity of the rating scale, modification of the structuring was made to provide for a more exact description of the performances. This was completed under the assumption that each category is not considered to be equal distance from the previous or subsequent categories. Making such changes will improve the ability and ease associated with the use of the model in future applications as well as its validity and reliability.

Table 3 provides the data that was taken into consideration when collapsing the rating scale structure. Frequency counts were investigated based on Linacre's (2002) recommendation of 10 uses per category. Any categories with less than 10 uses for certain items were collapsed in order to represent the best possible structure for those specific items and to avoid skewed distribution of item usage. Item 7 (Category 1), Item 8 (Category 4), Item 9 (Categories 1 and 4), Item 11 (Categories 1 and 4), Item 12 (Categories 1 and 4), Item 13 (Category 4), Item 14 (Category 1), Item 16 (Category 1), Item 17 (Category 1), Item 18 (Category 4), Item 19 (Category 1), Item 20 (Category 4), Item 21 (Category 4), Item 23 (Category 4), Item 24 (Category 4), Item 25 (Category 4), Item 26 (Category 4), Item 30 (Category 1), Item 31 (Category 1), Item 32 (Category 1), Item 35 (Categories 1 and 4), and Item 38 (Category 1) were collapsed into adjacent categories (based on frequency counts) in order to better serve the rating scale structure. Outfit mean squares (*MSE*) were examined for values ≥ 2.0 because such values would indicate excessive sporadic measures in the ratings. Items 21 and 25 (Category 4) were collapsed into adjacent categories in order to better serve the rating scale structure. Lastly, average observed logit measures were examined for violations of monotonicity. Monotonicity is the continuous advancement of step calibrations (Andrich, 1996). Agreement of monotonicity operates under the assumption that strongly agree is more difficult to endorse than agree and so forth. Therefore, if an item showed a violation in this monotonicity in the difficulty to endorse, the structure was collapsed. Item 10 was the only item that demonstrated violations of monotonicity and was therefore collapsed. Item 10 had already been discarded due to overfit, but if this was not the case, this would mean that only disagree or agree options were needed, as opposed to four separate Likert scale categories in the final rating scale. Without a qualitative investigation with the raters, it is hard to determine what might cause this result. Only an assumption can be made because an investigation such as this is outside of the scope of this study. In this study, the quantitative results of the rating scale category structure optimization based upon the analytics is the primary focus. Collapsing this item prior to developing the final rating scale would contribute to the usability of the final rubric in that raters would more easily be able to rate tempi by either disagreeing or agreeing that the tempi were appropriate for the style of the composition.

Table 3
Rating Scale Structure Analysis: Item Behavior of Category Usage, Average Observed and Expected Measures, and Outfit *MSE*

Item	Category Usage (%)				Average observed measure (Average expected measure)				Outfit MSE			
	1	2	3	4	1	2	3	4	1	2	3	4
1	8(8)	30(30)	48(48)	14(14)	–1.36(–1.14)	–.53(–.31)	.68(.60)	2.02(1.72)	0.80	0.70	0.60	0.80
2	4(4)	24(24)	58(58)	14(14)	–.32(–1.01)	.09(–.10)	.88(.85)	1.36(2.03)	1.90	1.30	0.90	1.50
3	4(4)	29(29)	57(57)	10(10)	–1.44(–1.16)	–.44(–.23)	.82(.76)	2.31(1.95)	0.80	0.70	0.70	0.80
4	11(11)	37(37)	45(45)	7(7)	–1.98(–1.53)	–.71(–.69)	.40(.30)	1.59(1.42)	0.60	0.70	0.80	0.90
5	13(13)	51(51)	31(31)	5(5)	–2.12(–1.78)	–.89(–.87)	.27(.18)	1.73(1.23)	0.80	0.80	0.80	0.60
†6	13(13)	32(32)	45(45)	10(10)	–1.65(–1.40)	–.57(–.60)	.46(.33)	1.08(1.43)	0.70	1.10	0.70	1.30
†7	6(6)	34(34)	49(49)	11(11)	–1.83(–1.21)	–.28(–.32)	.75(.64)	1.51(1.79)	0.60	1.00	0.90	1.20
†8	11(11)	40(40)	46(46)	3(3)	–2.14(–1.88)	–.82(–1.00)	–.10(.05)	2.04(1.19)	0.80	1.10	1.20	0.70
†9	9(9)	40(40)	44(44)	7(7)	–1.48(–1.5)	–.64(–.62)	.32(.38)	1.92(1.5)	1.10	1.10	0.90	0.80
10	5(5)	4(4)	63(63)	28(28)	.89(–.72)	–0.2*(0.1)	.88(.95)	1.97(2.12)	3.30	1.00	0.80	1.00
11	7(7)	35(35)	49(49)	9(9)	–.87(–1.33)	–.32(–.45)	.30(.53)	2.08(1.68)	1.80	1.20	1.10	1.00
12	5(5)	39(39)	47(47)	9(9)	–.82(–1.27)	–.51(–.34)	.74(.66)	1.86(1.81)	1.20	0.80	0.90	0.90
13	12(12)	37(37)	44(44)	7(7)	–1.54(–1.55)	–.78(–.71)	.29(.27)	1.54(1.38)	1.30	0.80	0.80	0.90
14	8(8)	32(32)	45(45)	15(15)	–1.06(–1.11)	–.11(–.28)	.53(.62)	1.57(1.72)	1.00	1.70	0.90	1.20
†15	6(6)	36(36)	52(52)	6(6)	–1.05(–1.47)	–.20(–.56)	.16(.47)	1.79(1.65)	1.30	1.60	1.60	0.90
†16	5(5)	15(15)	66(66)	14(14)	–1.06(–1.04)	–.08(–.18)	.78(.76)	1.78(1.98)	1.30	1.10	0.90	1.00
†17	4(4)	21(21)	59(59)	16(16)	–1.33(–.95)	–.02(–.05)	.89(.88)	2.09(2.06)	0.70	0.90	0.90	1.00
†18	10(10)	42(42)	42(42)	6(6)	–1.48(–1.60)	–.63(–.72)	.16(.30)	1.59(1.41)	1.10	1.20	1.10	0.90
†19	7(7)	26(26)	56(56)	11(11)	–1.18(–1.22)	.06(–.37)	.33(.57)	1.93(1.75)	1.30	1.70	1.10	0.90
20	17(17)	43(43)	35(35)	5(5)	–1.74(–1.80)	–1.08(–.95)	.13(.06)	1.67(1.12)	1.00	0.90	0.80	0.70
21	30(30)	32(32)	32(32)	6(6)	–1.97(–1.84)	–.98(–1.03)	–.13(–.10)	1.38(.89)	0.70	1.30	0.80	0.70
22	12(12)	35(35)	46(46)	7(7)	–1.79(–1.54)	–.90(–.71)	.38(.26)	1.94(1.39)	0.70	0.70	0.60	0.70

Item	Category Usage (%)				Average observed measure (Average expected measure)				Outfit MSE			
	1	2	3	4	1	2	3	4	1	2	3	4
23	31(31)	44(44)	24(24)	1(1)	–2.66(–2.61)	–1.78(–1.69)	–.42(–.61)	.72(.32)	1.10	0.80	0.70	0.90
24	22(22)	45(45)	28(28)	5(5)	–2.07(–1.89)	–.95(–1.02)	.12(–.01)	.36(.99)	0.80	1.10	0.70	2.00
†25	27(27)	41(41)	26(26)	6(6)	–2.02(–1.86)	–.78(–1.01)	.01(–.04)	–0.141	0.80	1.40	0.80	2.70
†26	28(28)	43(43)	24(24)	5(5)	–2.10(–1.96)	–.92(–1.09)	–.26(–.09)	1.03(.87)	0.80	1.20	1.00	1.10
†27	25(25)	30(30)	29(29)	16(16)	–1.50(–1.34)	–.42(–.59)	.45(.23)	.74(1.19)	0.70	1.50	0.60	1.80
†28	10(10)	26(26)	51(51)	13(13)	–1.32(–1.21)	–.30(–.42)	.39(.49)	1.88(1.62)	0.90	1.10	0.90	0.90
†29	31(31)	52(52)	12(12)	5(5)	–2.15(–2.08)	–.99(–1.13)	–.02(–.08)	–0.2673	0.90	1.10	1.10	2.80
30	7(7)	35(35)	47(47)	11(11)	–1.04(–1.24)	–.23(–.37)	.41(.59)	1.90(1.72)	1.30	1.30	1.00	0.90
31	5(5)	27(27)	55(55)	13(13)	–.98(–1.09)	–.10(–.20)	.69(.75)	1.93(1.92)	1.40	1.10	1.10	1.00
32	4(4)	24(24)	55(55)	17(17)	–.79(–.92)	–.23(–.02)	.99(.90)	2.04(2.06)	1.20	0.70	0.80	1.00
33	7(7)	37(37)	53(53)	3(3)	–2.29(–1.77)	–.95(–.86)	.29(.20)	2.23(1.40)	0.60	0.70	0.80	0.80
34	4(4)	27(27)	59(59)	10(10)	–1.61(–1.16)	–.41(–.23)	.80(.75)	2.32(1.96)	0.60	0.70	0.90	0.80
†35	2(2)	48(48)	42(42)	8(8)	–1.57(–1.14)	–.04(–.08)	.85(.97)	2.58(2.11)	0.80	1.00	1.10	0.70
†36	12(12)	42(42)	35(35)	11(11)	–1.64(–1.38)	–.48(–.53)	.34(.41)	1.75(1.46)	0.80	1.20	0.80	0.80
†37	15(15)	31(31)	43(43)	11(11)	–1.2(–1.39)	–.51(–.60)	.14(.31)	1.51(1.39)	1.40	1.30	1.00	0.90
†38	5(5)	27(27)	54(54)	14(14)	–.72(–1.05)	–.15(–.17)	.78(.77)	1.74(1.93)	1.40	1.	0.80	1.2

Note. Category 1 = strongly disagree; Category 2 = disagree; Category 3 = agree; Category 4 = strongly agree
†Category 1 = strongly agree; Category 2 = agree; Category 3 = disagree; Category 4 = strongly disagree
*Violation of monotonicity

A finalized version of the String Performance Rating Scale is shown in Appendix D. The final rating scale reflects the absence of discarded items as well as the modifications made to the structure of each item. These modifications were decided based on frequency of use, outfit measures, and monotonicity. The information provided in Table 3 and Appendix D can be used to answer the third research question.

Rubric development and defining performance criteria descriptors. Following the investigation of quantitative results for the rating scale, three content experts engaged in ex post facto qualitative analyses through the development of descriptors for each of the rating scale categories. This step was necessary in order to develop a rubric that would be considered useable in future string larger ensemble performance evaluation situations. The resulting rubric will allow for feedback to be shared with performers following the evaluation process and will be easier for raters to use in future evaluation settings.

The content experts provided the expertise needed to develop a rubric that reflected the appropriate wording and descriptions for each item in a way that would be meaningful to the middle school and high school string performance community. Careful consideration was taken to eliminate repetition and to ensure clarity and precision for each of the items in the rating scale. The original directionality within each item was removed in order to maintain a content-specific and non-directional learning outcome.

Preestablished anchors were used to develop statements for each criterion performance level descriptors (Vagias, 2006). Anchor selection included the categories of quality (Item 6), detraction (Items 7, 9, 17, 19, and 27), effectiveness (Items 8, 16, and 35), appropriateness (Item 11), frequency (Items 13, 14, 20, 21, 24, 26, 32, 37, and 38), problem (Items 18, 23, and 28), acceptability (Item 25), influence (Item 30), desirability (Item 31), and satisfaction (Item 36). The finalized rubric is shown in Figure 2.

CONCLUSIONS AND FUTURE RESEARCH

The first research question addresses how Rasch measurement analysis is used to reveal the psychometric quality (the validity and reliability) of the assessment used to evaluate large ensemble string performance. The item stems, raters, and performances were measured independent of one another, and the resulting metrics for each were used to determine which of each of the items, raters, and performances fit the model. This discernment between fit and misfit items, raters, and performances addresses the validity of the rating scale. Individual items that were considered to be misfit (outside of 0.80 and 1.20 logits) were discarded in favor of the items that adhered to the model.

The resulting rating scale is also considered to be reliable because a high reliability of separation is evident for persons, items, and performances. This reliability of separation ($Rel_{performances}$ = .97, Rel_{raters} = .98, and Rel_{items} = .93) increases the confidence that each of the measures accurately separates the facets that were measured. The high reliability of separation confirms that the characteristics were measured independent of one another within the context of the assessment.

Figure 2. Music performance rubric for string orchestra performance.

Tone Production				
6. Tone quality in varying registers	Tone quality is poor	Tone quality is fair	Tone quality is good	Tone quality is very good
7. Consistency of attacks	Unclear attacks always detract from performance	Unclear attacks sometimes detract from the performance	Unclear attacks never detract from the performance	
8. Tone while executing expressive gestures	The execution of expressive gestures has a major negative effect on tone quality	The execution of expressive gestures has a moderate negative effect on tone quality	The execution of expressive gestures does not have a negative effect on tone quality	
9. Consistency of tone across sections	Tone quality across sections detracts very much from the performance	Tone quality across sections detracts very little from the performance		
Rhythm and Pulse Accuracy				
11. Expressive pulse and tempo fluctuations	Expressive changes in tempo and pulse are inappropriate for the style	Expressive changes in tempo and pulse are appropriate for the style		
12. Sustained notes	Notes are not consistently held for full value	Notes are consistently held for full value		
13. Precision of attacks	Attacks are rarely executed with precision across the ensemble	Attacks are sometimes executed with precision across the ensemble	Attacks are consistently executed with precision across the ensemble	
14. Consistency of articulation	Rhythmic articulations are often inconsistent with the style of music and consistently lack ensemble uniformity	Rhythmic articulations are occasionally inconsistent with the style of music and sometimes lack ensemble uniformity	Rhythmic articulations are consistent with style of music and maintain ensemble uniformity	
16. Consistency of rhythmic stress	Rhythmic stress does not effectively communicate proper musical style	Rhythmic stress somewhat effectively communicates proper musical style	Rhythmic stress effectively communicates proper musical style	
17. Steadiness of pulse	A lack of steady pulse detracts much from the continuous flow of the music	Wavering steady pulse sometimes detracts from the continuous flow of the music	Control of steady pulse does not detract from the continuous flow of the music	
18. Appropriateness of tempo in technical passages	Tempo fluctuations during technical passages are a serious problem	Tempo fluctuations during technical passages are a moderate problem	Tempo fluctuations during technical passages are not at all a problem	
19. Subdivision of the rhythm	Inaccurate performance of subdivisions frequently detracts from solidly communicated tempo and meter	Inaccurate performance of subdivisions occasionally detracts from solidly communicated tempo and meter	Accurate performance of subdivisions contribute to solidly communicated tempo and meter	

Figure 2. *Continued.*

Intonation Accuracy				
20. Intonation of cadential points	Cadential points are rarely in tune		Cadential points are occasionally in tune	Cadential points are consistently in tune
21. Centered pitch	The pitch is rarely centered		The pitch is occasionally centered	The pitch is centered a great deal of the time
23. Overall intonation accuracy	Maintaining consistently good intonation in all registers is a serious problem during performance		Maintaining consistently good intonation in all registers is a moderate problem during performance	Maintaining consistently good intonation in all registers is not a problem during performance
24. Pitch adjustments	It is rarely evident that players are able to accurately and quickly adjust pitch when necessary		It is sometimes evident that players are able to accurately and quickly adjust pitch when necessary	It is frequently evident that players are able to accurately and quickly adjust pitch when necessary
25. Half step intonation	Half step intonation is unacceptable		Half step intonation is slightly unacceptable	Half step intonation is perfectly acceptable
26. Chromatic alterations intonation	Chromatic alterations are rarely in tune		Chromatic alterations are sometimes in tune	Chromatic alterations are consistently in tune
27. Presence of wrong notes	Wrong notes detract from the performance a great deal		Wrong notes occasionally detract from the performance	Wrong notes do not detract from the performance
28. Open string intonation	Out of tune open strings is a serious problem	Out of tune open strings is a moderate problem	Out of tune open strings is a minor problem	Out of tune open strings is not at all a problem
29. Intonation in technical passages	Intonation fluctuations during technical passages are a serious problem		Intonation fluctuations during technical passages are a moderate problem	Intonation fluctuations during technical passages are not at all a problem

Expressive Qualities/Stylistic Interpretation				
30. Presence of crescendo and diminuendo	Crescendo and diminuendo are not at all influential on effective expression		Crescendo and diminuendo are somewhat influential on effective expression	Crescendo and diminuendo are extremely influential on effective expression
31. Balance between melody and accompaniment	Balance between melody and accompaniment is undesirable		Balance between melody and accompaniment is desirable	Balance between melody and accompaniment is very desirable
32. Stylistically appropriate articulations	Stylistically appropriate articulations are never evident		Stylistically appropriate articulations are sometimes evident	Stylistically appropriate articulations are always evident
35. Connection of phrases	Ensemble does not meaningfully connect phrases		Ensemble meaningfully connects phrases	
36. Articulation	Articulations are inconsistent in passages with notes of a similar style, resulting in a very dissatisfactory performance	Articulations are often inconsistent in passages with notes of a similar style, resulting in a dissatisfactory performance	Articulations are sometimes inconsistent in passages with notes of a similar style, resulting in a satisfactory performance	Articulations are consistent in passages with notes of a similar style, resulting in a highly satisfactory performance
37. Contrast in dynamics	Dynamic contrasts are never evident	Dynamic contrasts are almost never evident	Dynamic contrasts are sometimes evident	Dynamic contrasts are frequently evident
38. Expressive modifications (>, sfz., rit., ten., cantabile)	Stylistic or expressive modifications are rarely appropriate or present in performance		Stylistic or expressive modifications are typically appropriate and somewhat present in performance	Stylistic or expressive modifications are appropriate and consistently present in performance.

The material was carefully examined in order to verify that the items accurately represented the construct that was being measured and any impeding material was omitted. Three content experts accounted for the validity of the items during initial reading, collection of materials, and discussion of the nature of the items. As a result of the psychometric analysis, 11 of the 38 items were considered misfit based on infit and outfit metrics. The 11 items that were removed only increase the future functionality of the measurement. Any items considered underfitting were too muted and did not provide enough variety in order to be considered valid, whereas the overfitting items were considered to be too sporadic and failed to adequately contribute to the scoring process. Items 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 30, 31, 32, 35, 36, 37, and 38 were considered to be valid because they fit the model (see Appendix D here: http://bcrme.press.illinois.edu/media/215/).

There are small standard errors associated with each rater and item that contributes to the degree in which the newly developed rating scale can be used for further investigation. This also contributes to evidence of strong precision. The precision and reliability evidence demonstrates that the measure was able to adequately distinguish between high and low performances while using the logit continuum. Further implications include the possibility of predicting the level of difficulty for items prior to data collection due to the fact that the sample is independent from the data to model fit. This particular aspect of the results could help in preparation efforts for ensemble rehearsals in the music classroom.

The second research question addresses how the items vary in difficulty, the raters in severity, and the ensembles in performance achievement. The variable map shows how each of these facets vary. In terms of items, the more difficult items to endorse are closer to the top and the easier items to endorse are at the bottom (Figure 1). Item 23 is considered to be the most difficult item to endorse ("ensemble performs with consistently good intonation in all registers"), and item 32 is considered to be the easiest item that fit the model ("stylistically appropriate articulations"). Item 10 cannot be considered the easiest item to endorse because the item was considered misfit (1.47 logits) and was therefore discarded.

Items that did not fit the model were considered invalid and were therefore discarded. Each of the four a priori categories contained item stems that were discarded. Five of the nine item stems from the tone production category were discarded, and each of those five item stems discarded were positively worded (Items 1, 2, 3, 4, and 5). Item 2 ("players use sufficient bow weight") was overfit, but the remaining discarded tone production items were underfit. This implies that raters might not be able to adequately describe tone production in a positive manner, or perhaps the raters have unrealistic expectations for what a characteristic tone might sound like in an exemplar string ensemble performance. Replication of this study might allow for more item stems relating to tone production to be tested in order to hopefully be able to provide more descriptors for this category in the final rating scale.

Two stems from the rhythm and pulse accuracy category were discarded (Items 10 and 15). Both of these items were considered overfit; Item 10 was positively worded and Item 15 was negatively worded. In contrast to the tone production category wherein 50% of the items were discarded, only 20% of the rhythm and pulse accuracy stems were discarded. The remaining items that were discarded were considered underfit. Two items from the pitch and intonation accuracy were discarded (Items 22 and 29). Item 22 was positively worded, and Item 29 was negatively worded. Two items were discarded from the expressive qualities/stylistic interpretation category (Items 33 and 34). Both of these items were positively worded. Aside from the five tone production item stems that were discarded, it seems as if the other stems that were discarded could perhaps be hard to hear and discern a true rating with the absence of a score during string performance.

Though modification is a possibility for future studies, the adaptation of the misfit items was outside of the scope of this study. For this reason, any items that did not fit the model were discarded and not included in the final rating scale or in the rubric. Future replications of the study would allow for the introduction of more stems in order to provide opportunities to further discover the unidimensionality and possible modification of such stems in order to counteract multidimensionality.

The third research question addresses the structure of the rating scale. The researchers investigated the null hypothesis that the original items share an identical response structure. This consideration provided the opportunity to show that the item stems do not all share the same structure. Inconsistencies in terms of monotonicity, frequency of use, and aspects of predictability provide evidence of the notion that item stems require different levels of answer responses according to the difficulty to endorse. The null hypothesis was rejected. The modifications that were made through collapsing the structure of certain item stems helped to improve the data to model fit. This collapsing also contributes to the overall usability of the rubric. For certain stems, it will suffice for future raters to either agree or disagree, not having to choose between strongly agree and strongly disagree for what are considered to be items that are more dichotomous in nature. Table 3 and Appendix D provide information that can be used to investigate the third research question.

The third research question also contributes to the overall usability of the final rubric. The changes in structure contributed to usability in that future raters will be able to better use the rubric in order to evaluate live performances of large string ensembles under specific time constraints that are typically characteristic of these situations. The overwhelming number of 38 original items, each with four levels of rating, was greatly reduced through the course of the study in order to create a valid, reliable, and more user-friendly rubric for the intended application.

The Rasch partial credit measurement model is ideal for this situation because the data can be treated differently in order to provide an accurate reading as to specific performance levels. Previous scoring practices operate under the assumption that rubric

data is considered to be interval-level data, but in this instance, the data received is actually ordinal-level data. The Rasch measurement model transforms the data from ordinal to interval data through metrics. Receiving data at a higher level means that the data is more meaningful in terms of how it relates to other and future performances. In developing this model, the data can be seen as counts that are directly correlated to a common measure as opposed to isolated measures that cannot be aligned with the overall relationship between the facets that were measured. This adjustment in data increases the possibility of the development of a validated assessment measure that should and can ideally be developed in relation to the predetermined standards for performance.

With its predictive nature, perhaps the most important future implication of a validated rating scale lies in its capability to be used in training and aligning those practices of individuals who rate performances. Rater severity measures can be used to help raters understand how performances should be evaluated in an effort to adequately align the opinions of those who serve as judges in such performance evaluation situations. The development of the rubric is an important process and can be revealing. However, future replications and uses of the scale itself can be used in order to help develop a more valid, reliable, and meaningful rating process for students, teachers, administrators, community members, and teacher candidates. This newly developed rubric can be used to provide valid and reliable empirical evidence of student performance achievement. Specifically, in a performance-based classroom setting that primarily utilizes rater-mediated assessments, this rubric provides a way for music educators to objectively rate and evaluate string ensemble performances.

As data-driven efforts continue to progress in the educational setting, the use of such rubrics in music performance evaluation will become a necessity. In moving forward, this particular rubric needs to be retested using a different population, including classroom teachers, in order to confirm that a distinction between expert raters and classroom teachers would yield positive results. As was evident in the pilot test of the Model Cornerstone Assessments, teachers evaluating their own student work and outside expert evaluators evaluating the same work did not demonstrate any form of differential rater functioning (i.e., bias) based upon their population grouping (Model Cornerstone Assessments, 2015). There is a high level of confidence that a continuation of the research and development of this rubric would yield the same positive results in future retesting. As such, a continued focus is needed to provide such performance assessment measures for the music education community.

This continued development of these assessment measures needs to involve the rigorous and concentrated efforts of content experts, expert raters, psychometric analysts, and influential political and community individuals as well. Once established and agreed upon, content experts and stakeholders could work to establish a valid and reliable assessment system that would provide achievement levels and descriptive forms of feedback for those performances that are either in need of improvement or should be commended for serving as an adequate model for future exemplary performance status.

The development of a valid rating scale also provides pedagogical speaking points for music teacher training programs. Specifically, the ranking of items in terms of difficulty presents the opportunity for preservice teachers to engage in meaningful conversations about the components of string performance and which items are going to be the most difficult. Helping teachers to discover and discuss the difficult aspects of rating string performance can help teacher candidates to formulate teaching strategies and approaches that will be effective. Furthermore, assuming that there is good data to model fit will provide the opportunity for stellar performances to be identified and will therefore supply preservice teachers and novice teachers with the ability to listen to exemplars as they learn how to achieve the same performance standard. Essentially, a valid and reliable string performance rating scale will help string music educators to better understand what distinguishes a great performance from a mediocre performance, therefore assisting teachers in being able to better align instruction in the classroom with valid and meaningful assessment processes.

AUTHORS' NOTE

This article was completed as part of a graduate study while enrolled at the University of Georgia during the 2016–17 academic year.

The authors presented material from this article at a research poster session at the Georgia Music Educators Association Convention in January 2016 as "Evaluation of a String Performance Rating Scale Using the Multifaceted Rasch Partial Credit Measurement Model." They also presented the material at the National Association for Music Education Research Conference in March 2016 as "Evaluation of a String Performance Rating Scale Using the Multifaceted Rasch Partial Credit Measurement Model."

SUPPLEMENTAL MATERIAL

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

REFERENCES

Andrich, D. (1996). Measurement criteria for choosing among models with graded responses. In A. V. Eye & C. C. Clogg (Eds.), *Categorical variables in developmental research* (pp. 3–35). San Diego, CA: Academic Press.

Asmus, E. P. (1999). Music assessment concepts. *Music Educators Journal, 86*(2), 19–24.

Blakeslee, M. (2004). Assembling the arts education jigsaw. *Arts Education Policy Review, 105*(4), 31–36.

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2014). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy, 29*(5), 711–745.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education, 39*(1), 52–71.

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

Colwell, R. (2003). The status of arts assessment: Examples from music. *Arts Education Policy Review, 105*(2), 11–18.

DeLuca, C., & Bolden, B. (2014). Music performance assessment: Exploring three approaches for quality rubric construction. *Music Educators Journal, 101*(1), 70–76.

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

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G., Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602.

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

Engelhard G., Jr. & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research and Perspectives, 9*(1), 40–45.

Hope, S., & Wait, M. (2013). Assessment on our own terms. *Arts Education Policy Review, 114*(1), 2–12.

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

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 86–106.

Linacre, J. M. (2010). More objections to the Rasch Model. *Rasch Measurement Transactions, 24*(3), 1298–1299.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34–43.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Model Cornerstone Assessments. (2015). *National Association for Music Education*. Retrieved from http://www.nafme.org/my-classroom/standards/mcas-information-on-taking-part-in-the-field-testing/

National Association for Music Education. (2015, August) Music model cornerstone assessment: performing ensemble proficient. Retrieved from https://nafme.org/wp-content/files/2014/11/Music_MCA_Ensemble_Performing_Proficient_2015–1.pdf

Parkes, K. (2010). Performance assessment: Lessons from performers. *International Journal of Teaching and Learning in Higher Education, 22*(1), 98–106.

Pellegrino, K., Conway, C. M., & Russell, J. A. (2015). Assessment in performance-based secondary music classes. *Music Educators Journal, 102*(1), 48–55.

Vagias, W. (2006). Likert-type scale response anchors. *Clemson International Institute for Tourism and Research Development, Department of Parks, Recreation and Tourism Management*. Clemson, SC: Clemson University.

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal, 98*(3), 36–42.

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

Wesolowski, B. C., 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, 5*, 662–678.

Whitcomb, R. (1999). Writing rubrics for the music classroom. *Music Educators Journal, 85*(6), 26–32.

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

Zaleski, D. J. (2014). An introduction to classroom assessment for today's music educator. *Illinois Music Educator, 75*(1), 58.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education, 50*(3), 245–255.