sempre:
Society for Education, Music and Psychology Research

Article

Research Studies in Music Education
1–25

DOI: 10.1177/1321103X18773103
journals.sagepub.com/home/rsm
SAGE

The psychometric evaluation of a wind band performance rubric using the Multifaceted Rasch Partial Credit Measurement Model

Andrew S. Edwards
The University of Georgia, USA

Kinsey E. Edwards
The University of Georgia, USA

Brian C. Wesolowski
The University of Georgia, USA

Abstract

The purpose of this study was to develop a valid and reliable rubric to be used for the evaluation of large ensemble wind band performances. The guiding questions for this study were: (a) what are the psychometric qualities (i.e., reliability and validity) of the scale developed to assess wind band ensemble performance at the high school level? (b) how do the items fit the model and vary in difficulty? (c) how does the structure of the rating scale vary across individual items? and (d) how can the rating scale be transferred into an informative rubric? The primary data analysis tool used in this study was the Multifaceted Rasch Partial Credit Measurement Model. Music content experts (N = 20) were solicited to evaluate 40 wind band performances, each evaluator listening to four. A 4-point Likert-type rating scale (e.g., Strongly Agree, Agree, Disagree, and Strongly Disagree) was used to evaluate each recorded performance. Results indicated good model data fit and resulted in a final rubric containing 24 items ranging from two to four performance categories. Implications for classroom teaching and consequential validity are discussed.

Keywords

assessment, invariant measurement, performance evaluation, Rasch, rubric, wind band

Decisions made by administrators, stakeholders, and policy makers regarding school programs are strongly guided by empirical student achievement data (Swan & Mazur, 2011; Wayman,

Corresponding author:
Andrew S. Edwards, Hugh Hodgson School of Music, The University of Georgia. 250 River Rd., Athens, GA 30602, USA.
Email: akedwards09@gmail.com

2005). Empirical evidence of student achievement is often at the foundation of policy-based decisions (O'Neal, 2012) and is traditionally derived from standardized testing results and stakeholders' contextualization of the results within specified content and performance standards (U.S. Department of Education, 2009). In the context of performance-based practices such as music and related performing arts, teachers are now held accountable for their role in student learning by providing evidence of student achievement using Student Learning Objective (SLOs) frameworks and other similar methods (Wesolowski, 2014; Wesolowski, Wind, & Engelhard, 2015). Many of these systems are created by either groups of local teachers without adequate training in measuring development processes or district leaders that do not have expertise in the content under evaluation (Wesolowski, 2012). When subject matter experts (e.g., practicing teachers) solely design these assessment systems, validity and reliability concerns exist due to the lack of psychometric analysis of the measurement instruments (Buckley & Marion, 2011). Conversely, when non-subject matter experts (e.g., district leaders or administrators) design these assessment systems, content and construct validity concerns exist due to their lack of content knowledge. Yet, administrators and other stakeholders are increasingly accepting these types of tests as a solidly defended option for providing empirical evidence of teacher effectiveness and student achievement, particularly in the area of music where the subject matter is often conceived of as being highly subjective. As a result, the gathered data often provides an inaccurate description of the actual teaching and learning that is occurring in the classroom. In these instances, decision-making bodies may potentially make decisions for educational programs based upon invalid data, posing unintended consequences stemming from the assessment process and related student achievement outcomes.

As noted by Wesolowski, (2012) the use of rubrics is a fruitful method for collecting empirical evidence of student achievement for music performance assessments. A rubric can be defined as "a set of scoring criteria used to determine the 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 performances in the future" (Asmus, 1999, p. 6). Rubrics are not only important for collecting student achievement data as a summative method of evaluation, but they are an essential formative assessment method to improve communication and expectations between teachers, students, and parents (Pellegrino, Conway, & Russell, 2015). When a student receives feedback in the form of a rubric, they not only have evidence of their performance achievement as a marked outcome, but they also have tangible information of what specific knowledge, skills, and abilities they need to improve in order to demonstrate achievement at a higher level. The use of rubrics improves the subjectivity of music assessment by using objective markers to evaluate student achievement in a clear and quantifiable way.

Professional development commonly focuses on the development of rubrics for the assessment of an individual student (Kan & Bulut, 2014; Sherman, 2006). However, little research has been done in the field of music on the implementation of rubrics for large ensemble performance evaluation; yet, ensemble performance is often the core activity for many music classrooms in secondary education (Wesolowski, Wind, & Engelhard, 2015). Without valid and reliable assessments being used to evaluate these performances, music students are not receiving the feedback necessary for improved learning. One common method for evaluating large group music performance assessments relies on a rating scale system of four to eight categories that are sum totaled for an overall achievement score (National Association for Music Education, 2016). There is commonly an attached set of qualitative comments from raters in either spoken or narrative form; however, this system lacks clear performance descriptors that specify the interpretation of each of the categories of the rating scale. These performance descriptors are the key pieces of evaluative information needed by judges to reduce construct-irrelevant

variability and key pieces of diagnostic information needed by teachers to better evaluate performance quality and set future instructional goals (DeLuca & Benjamin, 2014).

The purpose of this study was to develop a valid and reliable rubric for the evaluation of large ensemble wind band performance using psychometric principles of invariant measurement. This study was guided by the following research questions:

1. What are the psychometric qualities (i.e., reliability and validity) of the scale developed to assess wind band ensemble performance at the high school level?
2. How do the items fit the model and vary in difficulty?
3. How does the structure of the rating scale vary across individual items?
4. How can the rating scale be transferred into an informative rubric?

Background

A valid and reliable performance evaluation carries value far beyond informing policy-based decisions by administrators, stakeholders, and politicians. A measurement instrument that provides valid, reliable, and fair evaluative data is an important component to improving music teaching and learning throughout the school year. A formal, large ensemble performance evaluation traditionally occurs one time per year for most music students (Colwell, 1970). Outcomes of formal evaluations often guide teachers' decision-making processes for the next year by allowing the teacher insight into which pedagogical techniques that they have recently used were effective or ineffective (Banister, 1992). The use of an evaluation system that is valid and reliable from judge to judge and year to year can allow teachers to gauge the effectiveness of new classroom techniques, objectives, and repertoire (Abeles, Hoffer, & Klottman, 1994; Howard, 2002).

Colwell (1970) states that the evaluation of ensemble performances "can be the most meaningful evaluation of performance the student receives" (p. 105). Students are motivated by learning contexts that optimize success in learning (Austin, 1988). When evaluations are implemented in a valid, reliable, fair, and meaningful manner, they have the potential to hold great power in improving student motivation (Banister, 1992; Franklin, 1979; K. K. Howard, 1994; Hurst, 1994; Sweeney, 1998). Additionally, when formal evaluations are adjudicated using a statistically validated rubric, diagnostic feedback for improvement becomes more clear and meaningful. As such, students are then able to specifically identify how to improve their performances (Asmus, 1999).

Psychometric considerations

The primary methodology used for the development of measurement instruments in the field of music education has been factor analysis. Factor analysis studies for individual instrument scales include clarinet (Abeles, 1971), euphonium and tuba (Bergee, 1987), voice (Jones, 1986), snare drum (Nichols, 2005), auditioning vocalists (Pazitka-Munroe, 2003), guitar (Russell, 2010), and string instruments (Zdzinski & Barnes, 2002), among others. Musical ensemble performance scales have also been developed using factor analysis methods for choral ensembles (Cooksey, 1977), wind band ensembles (DeCamp, 1980), and string ensembles (Smith & Barnes, 2007). As outlined by Wesolowski (2017), however, the use of factor analysis can be limiting when raters mediate the measure development process.

In the context of performance assessments, data is derived through rater-mediation. More specifically, performances are estimated via a rater's interaction with evaluative cues set forth in a

measurement instrument (Engelhard, 2013). Variability in rater behavior can stem from each rater's unique experiences, background, and interaction with the measurement instrument (Wilson, 2005). In music, because raters are not trained to evaluate performances with machine-like consistency as in other high stakes performance assessment contexts such as writing (Wesolowski, Wind, & Engelhard, 2015), divergence of rater response is to be expected. In fact, divergence in judges' responses is often welcomed in music assessment as it provides an opportunity to improve performances from multiple expert perspectives (Wesolowski, 2012). However, when performances are being evaluated empirically, managing the quality of rater behavior is of upmost importance, as construct-irrelevant variability can skew the results (Wesolowski, Wind, & Engelhard, 2016). In this study, the family of Rasch Measurement Models was used to evaluate the quality of rater behavior in the scale development process. The benefit of the Rasch model when specifically using a rater parameter within the model is the five requirements for rater-invariant measurement (Engelhard, 2013). The five requirements for rater-invariant measurement include: (a) rater-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular raters that happen to be used for the measuring); (b) non-crossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from raters than a less able person; (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular persons used for calibration); (d) non-crossing rater response functions (i.e., any person must have a better chance of obtaining a higher rating from lenient raters than from more severe raters; and (e) variable map (i.e., persons and raters must be simultaneously located on a single underlying latent variable). When adequate fit of the Rasch model is observed, then the five requirements for invariant measurement as outlined by Engelhard (2013) are met. When the requirements of rater-invariant measurement are met within reasonable empirical boundaries, it is assumed the characteristics of performances, raters, and items do not create construct-irrelevant interference between the data and the model.

The Partial Credit version of the Rasch model (Masters, 1982) provides an important additional interaction parameter that allows for the analysis of the individual rating scale categories between each separate item in the measure. This provides the basis for challenging the basic assumption that rating scale categories are equidistantly spaced across all items. This also allows for an improvement in optimization and precision of the item scales being analyzed due to the review of the unique step difficulties for each rating scale category for each respective item. The statistical output of the Rasch model analysis additionally allows for the review of category distribution within items and appropriate discrimination between performances (Linacre, 2002). The Rasch family of models provides not only a meaningful method for the development of valid and reliable measures, but provides a more consistent method for evaluating rater quality within the context of performance-based assessments. Analysis of the rating data in this study was performed using the computer program *FACETS* (Linacre, 2014).

Method

Rater content experts

Raters ($N = 20$) solicited for this study were music educators with an average of 14.2 years ($SD = 10.3$) of secondary music teaching experience in one southern state in the United States. The raters all had successful experiences directing ensembles that participate annually in large ensemble or festival-style performances. None of the raters had any influence on the development of the items in the rating scale or knowledge of the ensembles for which they were providing ratings.

Development of a priori item pool

The items used in the initial rating scale were developed from both the initial item pool used in DeCamp's (1980) study as well as through a collaboration of three subject matter experts. DeCamp's (1980) study employed a factor analysis approach to high school wind band ratings. This study resulted in 117 potential item stems that were then combined and aggregated into 39 initial items used in this study. The items were divided into four domains based on the National Association for Music Education (NAfME) Model Cornerstone Assessments (MCAs) for ensemble performance: (a) tone production ($n = 8$); (b) rhythm and pulse ($n = 10$); (c) pitch and intonation ($n = 7$); and (d) expressive qualities ($n = 14$) (National Association for Music Education, 2015). Eighteen of the item stems were presented as a positive statement and 21 were presented as negative statements as confirmed with 100% agreement by three subject matter experts not involved in the evaluation process. The items were presented in randomized order for each performance in order to defend against rater fatigue and to attempt to minimize rater errors such as central tendency, halo effect, and/or response sets during the evaluation process (Berk, 2010). A Likert-type rating scale structure was developed for each item. Each item stem was followed by four Likert-type response categories: *Strongly Agree, Agree, Strongly Disagree*, and *Disagree*. The rater responses were collected using an online Google form for final analysis preparation. The original rating scale is provided in Figure 1.

1. Tone is well controlled at all volume levels	SD D A SA
2. Characteristic tone is used throughout performance	SD D A SA
3. Tone quality is consistently rich in all registers	SD D A SA
4. Characteristic balance is achieved between scored voices	SD D A SA
5. Tone is compromised while executing expressive gestures	SD D A SA
6. Inconsistent tone across sections	SD D A SA
7. Tone is harsh due to overblowing	SD D A SA
8. Tone lacks proper air support	SD D A SA
9. Tempo fluctuations are stylistically characteristic	SD D A SA
10. Attacks are precise	SD D A SA
11. Tempo is appropriate for selection	SD D A SA
12. Note changes are smooth and even	SD D A SA
13. Rhythms are accurately articulated across the ensemble	SD D A SA
14. Tempo fluctuates during technical passages	SD D A SA
15. Rhythmic stress of strong and weak beats are uncharacteristic for style	SD D A SA
16. Steady pulse is unclear in performance	SD D A SA
17. Ensemble releases are inconsistent	SD D A SA

Figure 1. (Continued)

18. Rhythmic figures are subdivided inaccurately	SD D A SA
19. Players are able to accurately and quickly adjust pitch when necessary	SD D A SA
20. Intonation is accurate and consistent during crescendos and diminuendos	SD D A SA
21. Key signatures and key changes are accurately performed	SD D A SA
22. Technical deficiency limits quality of performance	SD D A SA
23. Harmonic intonation detracts from performance	SD D A SA
24. Intonation inaccuracy throughout the performance	SD D A SA
25. Performance contains incorrect pitches	SD D A SA
26. Performance maintains good attention to dynamics	SD D A SA
27. Proper balance between melody and accompaniment	SD D A SA
28. Stylistically appropriate articulations	SD D A SA
29. Displays effective musical communication	SD D A SA
30. Appropriate inflection at cadential points	SD D A SA
31. Inconsistent connection of phrases	SD D A SA
32. Articulations lack consistency in performance	SD D A SA
33. Crescendo and diminuendo are poorly graduated	SD D A SA
34. Performance is choppy with little consideration for phrasing	SD D A SA
35. Stylistic or expressive modifications (such as >, sfz, rit., ten., cantabile) are not evident in performance	SD D A SA
36. Appropriate style is effectively communicated through performance	SD D A SA
37. Some individuals disrupt the balance of the ensemble	SD D A SA
38. Lack of clarity in articulation	SD D A SA
39. Technical demands detract from balance of ensemble	SD D A SA

Figure 1. Original rating scale.

Performance stimuli

All performance recordings were collected from one formal large group performance evaluation during the 2012 concert festival season. The performances were all professionally recorded. More specifically, the same sound engineer, using the same equipment, and in the same concert hall recorded all of the recordings over a three-day festival event. Seventy-four performances were originally collected from 25 ensembles. The names of the musical selections and performers were made anonymous. An online random number generator was then used in order to assign the performance recordings to 20 distinct raters. Use of the performance recordings and participation of the raters in this study were permissible through the respective research institution's ethics and review board.

Evaluation collection and rater assessment network

Performance recordings were distributed to the raters using the online Dropbox service. Raters had access only to their four a priori assigned performances. The raters then used a Google Form to complete the evaluation for each individual performance. The evaluation process was designed as an incomplete assessment network (Engelhard, 1997). In this process each rater evaluated four performances, with two overlapping performances per rater (e.g., Rater 1 evaluated performances 1, 2, 3, and 4, Rater 2 evaluated performances 3, 4, 5, and 6, etc.). The final rater was linked to the first rater (i.e., final rater evaluated performances 39, 40, 1, 2.) whereby connectivity is achieved in the assessment network, allowing for the direct comparison of raters across all performances. This specific network design was found to provide the best model fit and smallest standard errors for an incomplete assessment network (Wind, Engelhard, & Wesolowski, 2016). When the evaluations were completed and all data was collected, the negatively worded items were reverse coded before undergoing empirical analysis in order to establish similar directionality in data analysis.

Results

Variable map

The variable map (see Figure 2) is a graphical representation of the analyzed data that operationally defines the latent construct of this study (i.e., performance achievement of high school wind bands). The variable map displays the three facets that were used in the model: (a) item difficulty; (b) performance ability; and (c) rater severity. The logit scale, found in the first column, can be considered the "ruler" for comparative measurement of each of the facets. The placement on the logit scale represents data that is interval level in nature, enabling one to make direct comparisons between data across the separate facets. The second column displays each of the performance measures in order from highest achieving at the top of the column to lowest achieving at the bottom of the column. Each individual performance is represented by a single asterisk on the variable map. The measures ranged from 2.87 to –2.43 logits (M =– .03, SD = 1.22, N = 40). The third column of the variable map displays the severity measurements of the raters from most severe at the top of the column to least severe at the bottom. The measures ranged from 2.05 to −1.76 logits (M = .00, SD = .97, N = 20). The fourth column displays the item difficulty measurement from most difficult near the top of the column to least difficult near the bottom. The measures ranged from 1.24 to −1.43 (M = .00, SD = .75, N = 39).

Summary statistics

Summary statistics are provided in Table 1. The data represented in this table indicates the overall significant difference between performances, $\chi^2 = 1516.6$, $p < .01$, raters, $\chi^2 = 976.7$, $p < .01$, and items, $\chi^2 = 606.2$, $p < .01$. The reliability of separation statistic can be conceptualized as the sensitivity of the measurement instrument to distinguish, or separate, individual elements within a particular facet and its ability to reproduce the logit locations. The reliability of separation for performances (Rel = .98), raters (Rel = .98), and items (Rel = .94) provides confidence to imply that there is enough separation to confirm the construct validity of the measurement instrument. The mean square fit (MSE) statistics (Infit MSE = .99 and Outfit MSE = 1.01) are close to the expected value of 1.00, demonstrating reasonably good data fit to the

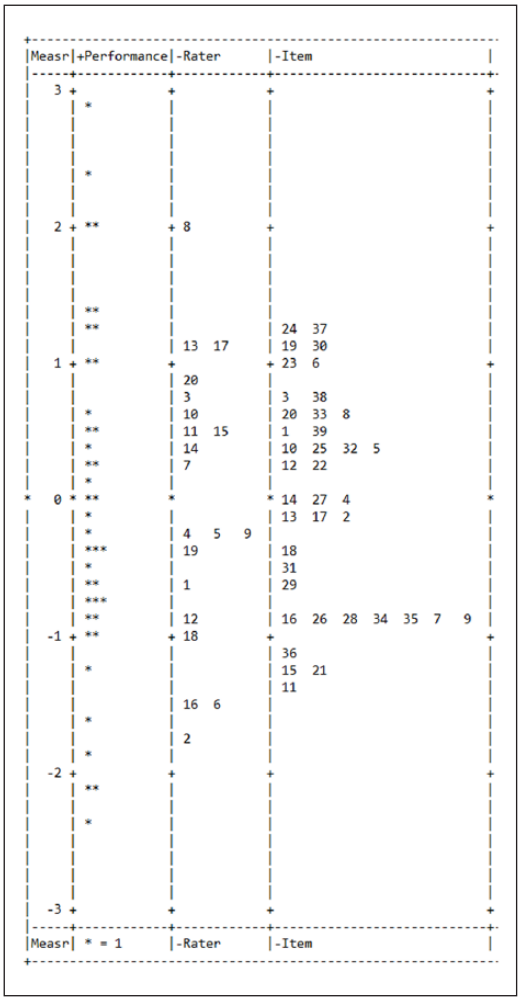

Figure 2. Variable map.

model based upon Wright and Linacre's (1994) acceptable range for parameter-level mean square statistics (0.50 to 1.50).

Calibration of performances

Table 2 provides a detailed report of the analysis of performances. The *measure* statistic indicates the logit scale location of the performance's achievement. A higher measure number indicates greater performance achievement and a lower measure indicates a lower performance achievement. The highest achieving performance was performance 20 (2.87 logits), and the lowest achieving performance was performance 11 (−2.43 logits). The infit *MSE* statistic was used to determine evidence of performance misfit. Wright and Linacre (1994) and Engelhard (2009) indicate that the best range for data misfit at the element level is between .80 and 1.20 logits for high stakes assessments. Performances demonstrating misfit due to over-fit (fit indices

Table 1. Summary statistics from the PC-MFR model.

	Facets		
	Performance (θ)	Rater (λ)	Item (δ)
Measure (Logits)			
Mean	−.03	0.00	0.00
SD	1.23	0.98	0.78
N	40	20	39
Infit *MSE*			
Mean	0.99	0.99	0.99
SD	0.33	0.30	0.21
Std. Infit *MSE*			
Mean	−0.20	−0.30	−0.10
SD	2.10	2.80	1.40
Outfit *MSE*			
Mean	1.01	1.01	1.01
SD	0.34	0.30	0.23
Std. Outfit *MSE*			
Mean	−0.10	−0.10	0.0
SD	2.10	2.80	1.50
Separation statistics			
Reliability of Separation	0.98	0.98	0.94
Chi-Square	1516.6*	976.7*	606.2*
Degrees of Freedom	39	19	38

*$p < .01$.

greater than 1.20 logits) include performances 4, 6, 7, 14, 18, 25, 35, 39, and 40. Performances demonstrating misfit due to under-fit (fit indices less than .80 logits) include performances 3, 9, 12, 22, 23, 24, 30, 31, 32, 33, 34, and 38.

Calibration of raters

Table 3 provides a detailed data analysis of raters. The measure indicates the logit scale location of the rater severity. A higher measure number indicates greater rater severity and a lower number demonstrates lower rater severity (i.e., leniency). The most severe rater was rater 8 (observed average = 2.13, logit measure = 2.05) and the least severe rater was rater 2 (observed average = 3.19, logit measure = −1.76). Raters 6, 12, 14, and 20 demonstrated misfit as evidenced by Infit *MSE* scores greater than 1.20. This substantively indicates ratings that were too sporadic (i.e., too unpredictable) to fit the model. Raters 3, 4, 17, and 19 demonstrated misfit due to an Infit *MSE* score less than .80. This substantively indicates ratings that were too muted (i.e., too predictable) to fit the model.

Calibration of items

Table 4 provides a detailed data analysis of rating scale items. The measure indicates the logit scale location of the difficulty of each item. A higher logit measure number indicates a more difficult item and a lower logit measure indicates a less difficulty (i.e., easier) item. The most

Table 2. Calibration of performance facet.

Performance number	Observed average	Measure	Standard error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
20	3.21	2.87	0.20	0.83	−1.00	0.96	−0.20
33	3.35	2.32	0.22	0.74	−1.60	0.72	−1.60
14	2.99	2.05	0.19	1.98	4.90	2.00	5.00
17	3.35	2.03	0.22	1.16	0.90	1.29	1.50
25	3.24	1.43	0.20	1.32	1.90	1.22	1.30
6	3.30	1.39	0.20	1.43	2.50	1.52	3.00
27	2.60	1.31	0.19	1.01	0.10	1.00	0.00
28	3.18	1.22	0.20	0.92	−0.40	0.97	−0.10
2	2.79	1.02	0.19	1.04	0.20	1.14	0.80
21	3.04	1.00	0.20	0.81	−1.20	0.83	−1.00
36	2.69	0.62	0.19	1.19	1.10	1.17	1.00
31	2.77	0.46	0.19	0.49	−3.70	0.50	3.60
39	2.42	0.44	0.19	1.21	1.30	1.24	1.40
29	2.83	0.36	0.19	1.07	0.40	1.08	0.50
24	2.54	0.29	0.19	0.70	−2.00	0.71	−1.90
13	2.32	0.28	0.19	0.84	−1.00	0.82	−1.10
15	2.51	0.10	0.19	0.85	−0.90	0.89	−0.60
40	2.59	−0.01	0.19	1.41	2.20	1.36	2.00
30	2.94	−0.03	0.19	0.56	−3.20	0.55	−3.30
4	2.43	−0.12	0.19	1.28	1.60	1.29	1.70
9	2.35	−0.24	0.19	0.53	−3.50	0.53	−3.50
10	2.49	−0.34	0.19	0.86	−0.80	0.87	−0.70
8	2.32	−0.41	0.19	1.16	1.00	1.11	0.70
7	2.18	−0.42	0.19	1.21	1.30	1.19	1.20
18	2.31	−0.46	0.19	1.51	2.80	1.54	2.90
1	2.17	−0.59	0.19	0.92	−0.40	0.92	−0.40
26	1.91	−0.67	0.20	1.03	0.20	1.11	0.60
19	2.22	−0.69	0.19	0.87	−0.80	0.85	−0.90
5	2.18	−0.72	0.19	1.09	0.60	1.06	0.40
34	2.23	−0.78	0.19	0.65	−2.50	0.67	−2.30
12	2.31	−0.81	0.19	0.72	−1.90	0.73	−1.80
38	2.19	−0.89	0.19	0.54	−3.40	0.57	−3.20
35	2.48	−0.98	0.19	1.70	3.60	1.78	3.90
23	2.41	−1.01	0.19	0.50	−3.70	0.53	−3.50
32	2.27	−1.24	0.19	0.77	−1.50	0.76	−1.60
3	1.82	−1.59	0.20	0.75	−1.80	0.76	−1.60
37	2.12	−1.83	0.19	1.02	0.10	1.04	0.20
16	1.62	−2.08	0.21	1.10	0.60	1.08	−0.50
22	2.19	−2.13	0.19	0.71	−2.00	0.71	−2.00
11	1.74	−2.43	0.20	1.19	1.10	1.41	2.20
Mean	2.51	0.00	0.19	0.99	−0.10	1.01	0.00
SD	0.28	0.78	0.01	0.21	1.40	0.23	1.50

Note. Presented in measure order from highest achievement to lowest achievement.

Table 3. Calibration of rater facet.

Rater number	Observed average	Measure	Standard error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
8	2.13	2.05	0.14	0.88	−1.10	0.89	−0.90
13	2.34	1.15	0.13	1.04	0.30	1.05	0.50
17	2.14	1.07	0.14	0.44	−6.70	0.44	−6.70
20	2.24	0.93	0.13	1.29	2.40	1.34	2.70
3	2.53	0.79	0.13	0.72	−2.70	0.73	−2.60
10	1.98	0.67	0.14	1.15	1.40	1.13	1.10
11	1.94	0.48	0.14	1.14	1.20	1.22	1.70
15	2.10	0.45	0.14	0.81	−1.80	0.80	−1.90
14	2.42	0.38	0.14	1.59	4.50	1.61	4.70
7	2.47	0.22	0.14	1.18	1.60	1.19	1.60
5	2.64	−0.19	0.14	0.91	−0.70	0.95	−0.40
4	2.63	−0.22	0.13	0.53	−4.80	0.55	−4.60
9	2.78	−0.26	0.14	0.90	−0.90	0.92	−0.60
19	2.71	−0.41	0.14	0.66	−3.40	0.69	−3.00
1	2.58	−0.61	0.14	1.05	0.40	1.02	0.10
12	2.61	−0.83	0.14	1.27	2.10	1.27	2.20
18	2.74	−0.94	0.14	0.81	−1.70	0.86	−1.20
16	3.07	−1.46	0.14	1.05	0.40	1.02	0.10
6	3.05	−1.51	0.14	1.60	4.60	1.59	4.60
2	3.19	−1.76	0.15	0.85	−1.30	0.95	−0.30
Mean	2.51	0.00	0.14	0.99	−0.30	1.01	−0.10
SD	0.35	0.98	0.00	0.30	2.80	0.30	2.80

Note. Presented in measure order from most severe to least severe.

difficult item was item 24 (*Intonation inaccuracy throughout the performance*; observed average = 1.92, logit measure = 1.24). The easiest item was item 11 (*Tempo is appropriate for selection*; observed average = 3.05, logit measure = −1.43). Items 1, 2, 5, 10, 12, 28, 29, 33, 34, and 38 demonstrate misfit due to an Infit *MSE* statistic less than .80, which demonstrates items that are too predictable when rating performances. Items 6, 11, 19, 23, and 25 demonstrate misfit due to an Infit *MSE* statistic above 1.20, which demonstrates items that are too unpredictable at distinguishing between higher achieving and lower achieving performances. Items demonstrating misfit were removed from the final rubric.

Rating scale category diagnostics

As the data from the facet calibrations began to help shape the final rating scale, a detailed analysis of the function of each item was performed as a result of employing the partial credit (PC) version of the model. A rating scale structure cannot be complete until, as Linacre (2002) suggests, careful steps toward optimization of the rating scale category structures are taken. The four-category Likert-type rating scale construction was carefully modified based on individual item diagnostic statistics in order to address and improve issues of construct validity that are associated with the rating scale as a whole. Analysis of the rating scale structure is what Bond and Fox (2007) claim is vital in order to clarify the meaning of and improve the usability of the measure. The item behavior statistics can be viewed in Table 5. This table defines the

Table 4. Calibration of item facet.

Item number	Observed average	Measure	Standard error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
24	1.92	1.24	0.18	1.20	1.20	1.24	1.40
37	2.05	1.22	0.19	0.92	−0.50	0.91	−0.50
30	2.70	1.15	0.23	1.02	0.10	1.00	0.00
19	2.16	1.09	0.20	1.42	2.40	1.43	2.50
6	2.14	1.04	0.19	1.42	2.40	1.47	2.80
23	2.08	0.96	0.19	1.30	1.70	1.26	1.60
38	2.20	0.73	0.19	0.76	−1.60	0.77	−1.60
3	2.20	0.70	0.18	0.94	−0.30	0.96	−0.20
20	2.24	0.68	0.20	1.18	1.10	1.17	1.00
33	2.78	0.66	0.21	0.74	−1.80	0.81	−1.20
8	2.27	0.61	0.19	0.92	−0.40	0.92	−0.40
39	2.28	0.49	0.17	1.05	0.30	1.17	1.10
1	2.31	0.45	0.19	0.64	−2.70	0.64	−2.70
10	2.35	0.40	0.20	0.71	−2.00	0.71	−2.00
5	2.35	0.37	0.19	0.69	−2.20	0.70	−2.20
25	2.40	0.36	0.18	1.37	2.20	1.42	2.40
32	2.33	0.32	0.19	1.17	1.00	1.19	1.10
22	2.33	0.27	0.16	1.00	0.00	1.32	1.80
12	2.47	0.24	0.21	0.77	−1.50	0.74	−1.70
14	2.51	0.04	0.18	1.07	0.50	1.10	0.60
4	2.51	0.02	0.19	1.14	0.90	1.12	0.80
27	2.56	0.01	0.18	0.90	−0.60	0.87	−0.70
13	2.53	−0.13	0.19	0.90	−0.60	0.89	−0.70
2	2.55	−0.16	0.18	0.73	−1.90	0.72	−2.00
17	2.49	−0.17	0.20	0.94	−0.30	0.94	−0.30
18	2.62	−0.34	0.18	1.14	0.80	1.27	1.60
31	2.58	−0.50	0.21	0.87	−0.80	0.85	−0.90
29	2.78	−0.64	0.18	0.77	−1.50	0.73	−1.70
7	2.85	−0.81	0.20	1.02	0.10	1.04	0.20
9	2.84	−0.83	0.18	0.99	0.00	1.09	0.60
28	2.76	−0.86	0.21	0.74	−1.70	0.72	−1.90
16	2.84	−0.89	0.19	0.97	−0.10	0.96	−0.10
34	2.80	−0.89	0.19	0.76	−1.60	0.76	−1.60
26	2.83	−0.89	0.20	0.94	−0.30	1.01	0.00
35	2.78	−0.91	0.20	1.03	0.20	1.06	0.40
36	2.84	−1.07	0.20	0.84	−1.00	0.82	−1.10
15	2.94	−1.24	0.20	1.16	1.00	1.15	0.90
21	2.89	−1.30	0.22	1.11	0.60	1.07	0.40
11	3.05	−1.43	0.21	1.34	1.80	1.42	2.10
Mean	2.51	0.00	0.19	0.99	−0.10	1.01	0.00
SD	0.28	0.78	0.01	0.21	1.40	0.23	1.50

Note. Presented in measure order from most difficult to least difficult.

empirical data that was used to determine final construction of the scale in its reverse coded form to preserve directional consistency for positively and negatively worded items.

Table 5. 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
3	18 (23)	34 (43)	22 (28)	–	−2.04 (−2.07)	−1.05 (−1.03)	.11 (.18)	2.00 (1.66)	1.00	1.00	1.00	0.70
4	8 (10)	30 (38)	35 (44)	–	−1.53 (−1.7)	−.83 (−.76)	.65 (.49)	1.45 (2.12)	1.10	0.80	1.00	1.70
†7	–	16 (20)	47 (59)	12 (15)	−1.01 (−1.14)	−.27 (−.34)	.79 (.85)	2.61 (2.54)	0.90	1.30	0.90	1.00
†8	13 (16)	37 (47)	24 (30)	–	−2.04 (−2.09)	−1.15 (−1.02)	.45 (.26)	1.71 (1.81)	1.10	0.90	0.80	1.00
9	–	20 (25)	38 (48)	17 (21)	−.52 (−1.04)	−.42 (−.23)	.88 (.88)	2.45 (2.38)	1.90	0.80	1.10	0.90
13	–	33 (41)	31 (39)	9 (11)	−1.26 (−1.56)	−.86 (−.60)	.84 (.62)	2.13 (2.17)	1.20	0.60	0.60	1.30
†14	10 (13)	27 (34)	35 (44)	8 (10)	−1.15 (−1.68)	−.98 (−.76)	.39 (.44)	2.32 (2.03)	1.80	0.90	0.90	0.80
†15	–	18 (23)	42 (53)	17 (22)	−.79 (−.77)	.21* (.04)	.96 (1.21)	3.00 (2.76)	2.30	1.40	0.80	0.80
†16	–	21 (26)	39 (49)	16 (20)	−1.59 (−1.01)	−.11 (−.20)	1.01 (.95)	2.35 (2.47)	0.60	1.10	0.90	1.20
†17	–	40 (50)	29 (6)	–	−1.22 (−1.56)	−.55 (−.54)	.56 (.77)	3.03 (2.37)	1.10	1.20	0.90	0.50
†18	8 (10)	–	34 (44)	11 (14)	−.65 (−1.40)	0.3672	.40 (.57)	2.46 (2.16)	2.70	0.90	0.80	0.80
20	11 (14)	43 (54)	22 (28)	–	−2.17 (−2.18)	−.96 (−1.09)	.07 (.28)	1.61 (1.85)	1.00	1.20	1.30	1.10
21	–	19 (24)	48 (60)	12 (15)	−.09 (−.72)	.37 (.09)	1.16 (1.34)	3.27 (3.04)	1.20	1.30	0.90	0.90
†22	20 (25)	27 (34)	20 (25)	13 (16)	−1.50 (−1.64)	−.74 (−.69)	.19 (.32)	1.72 (1.63)	1.00	1.90	1.50	1.10
†24	28 (35)	35 (44)	12 (15)	–	−2.29 (−2.36)	−1.20 (−1.23)	−.07 (.01)	.99 (1.37)	1.00	1.70	1.00	1.30
26	–	21 (26)	43 (54)	13 (16)	1.35 (−1.04)	0.17 (.01)	1.03 (.99)	2.88 (2.60)	3.60	0.40	0.50	0.80
27	11 (14)	21 (26)	40 (50)	8 (10)	−1.64 (−1.67)	−.98 (−.79)	.44 (.39)	2.23 (2.02)	1.00	0.70	0.80	0.90
30	–	31 (39)	42 (53)	–	–	−2.21 (−2.17)	−.74 (−.82)	.67 (.93)	–	1.00	1.00	1.10
†31	–	37 (46)	34 (43)	–	−.63 (−1.32)	−.48 (−.33)	1.05 (1.00)	2.96 (2.65)	1.20	0.80	0.80	0.80
†32	10 (13)	41 (51)	22 (28)	–	−1.8 (−1.86)	−.77 (−.81)	−.52 (.45)	1.45 (1.96)	1.00	1.20	0.90	1.90
†35	–	25 (31)	42 (53)	11 (14)	−.97 (−1.02)	−.04 (−.15)	.98 (1.09)	2.88 (2.74)	1.00	1.30	1.10	0.80
36	–	23 (29)	41 (51)	14 (18)	−.94 (−.88)	−.16 (−.04)	1.13 (1.16)	3.01 (2.74)	0.90	0.80	0.80	0.80
†37	22 (28)	35 (44)	20 (25)		−2.32 (−2.48)	−1.72 (−1.38)	.45 (−.07)	.79 (1.44)	1.30	0.70	0.50	1.30
†39	18 (23)	30 (38)	24 (30)	8 (10)	−1.58 (−1.89)	−1.08 (−.90)	.14 (.24)	1.98 (1.69)	1.20	1.50	1.10	0.70

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.

The analysis of the rating scale structure included four stages. The first stage of analysis included category eliminations based upon the analysis of category frequency counts for each item of the original scale. Linacre's (2002) recommendation for this process is 10 uses per category. However, due to the smaller scale of the study, 10% of total responses per item (8 uses) was used as an acceptable cut point for each category. This resulted in the collapsing of adjacent categories in 18 items in order to guard against irregular skew and distributions. These categories include: item 3 (category 4), item 7 (category 1), item 8 (category 4), item 9 (category 1), item 13 (categories 1), item 15 (category 1), item 16 (category 1), item 17 (categories 1 and 4), item 20 (category 4), item 21 (category 1), item 24 (category 4), item 26 (category 1), item 30 (categories 1 and 4), item 31 (categories 1 and 4), item 32 (category 4), item 35 (category 1), item 36 (category 1), and item 37 (category 4).

The second stage of analysis included analysis of the Outfit *MSE* statistics for each of the categories. Values greater than or equal to 2.00 indicate excessive unpredictability in the rating categories that were being used unexpectedly by raters (Linacre, 2002). After eliminating items from misfit and categorical frequency the only item found exceeding the value of 2.00 was item 18 (category 2). This category was therefore collapsed with its adjacent category, category 1.

The third stage of analysis included analysis of monotonicity. Monotonicity refers to the unidirectional movement up the logit scale as the categories of an item increase in difficulty (i.e., proper step ordering). Category 2 of items 15 and 26 was collapsed with its adjacent categories as a result of violations of monotonicity (e.g.. Item 15, category 1 shows a logit scale rating of .77 while category 2 shows a logit rating of .21 which is less than .77, violating the unidirectional ordering of categories). This continuous advancement of step calibration within the measurement tool is vital to strong construct validity (Andrich, 1996).

The final phase of category diagnostics involved reviewing the remaining categories in order to show large enough measure difference between categories. The threshold of .70 logits was used as the difference in observed average measures to ensure a proper separation of categories. Categories 1 and 2 of item 14 revealed a difference of .17 logits. Categories 1 and 2 of item 39 revealed a difference of .50 logits. As a result, it made substantive sense to collapse the two categories due to a lack of clear definition and separability between categories. The resulting scale with remaining items and collapsed categories is provided in Figure 3, which demonstrates the adjustments made using Linacre's guidelines for optimizing rating scale structure.

Rubric design

Following the finalization of the wind band rating scale, the results were analyzed and reviewed in order to develop a rubric that would better align with the empirical results of the rating scale. This rubric was developed to provide specific category descriptors that would allow all scale users to be able to clearly understand the resulting performance evaluations and provide information for improvement in the future. Three content experts reviewed the rubric for content and face validity. The rubric was formed using Vagias' (2006) anchors from the categories of Affect (item 15), Appropriateness (item 9), Barrier (items 22, 37, and 39), Dichotomous (items 30 and 31), Frequency (items 7, 8, 9, 13, 14, 17, 18, 20, 21, 26, 32, 35, and 36), and Problem (items 3, 4, 5, 23, and 27). The final rubric is provided in Figure 4. The rubric should be viewed as a tool for feedback in order to guide future study by the ensemble. The rubric reveals basic categories of musicianship and the anchors that describe the ensemble's ability within that category. The anchor statements were developed in close relation to the statements in the rating scale to increase the reliability of feedback when the rubric is used in coordination with the rating scale.

3. Tone quality is consistently rich in all registers	Strongly Disagree	Disagree	Agree
4. Characteristic balance is achieved between scored voices	Strongly Disagree	Disagree	Agree
7. Tone is harsh due to overblowing	Strongly Disagree	Disagree	Agree
8. Tone lacks proper air support	Disagree	Agree	Strongly Agree
9. Tempo fluctuations are stylistically characteristic	Disagree	Agree	Strongly Agree
13. Rhythms are accurately articulated across the ensemble	Disagree	Agree	Strongly Agree
14. Tempo fluctuates during technical passages	Strongly Disagree	Disagree	Agree
15. Rhythmic stress of strong and weak beats are uncharacteristic for style	Strongly Disagree	Disagree	Agree
16. Steady pulse is unclear in performance	Strongly Disagree	Disagree	Agree
17. Ensemble releases are inconsistent	Strongly Disagree	Disagree	Agree
18. Rhythmic figures are subdivided inaccurately	Strongly Disagree	Disagree	Agree
20. Intonation is accurate and consistent during crescendos and diminuendos	Strongly Disagree	Disagree	Agree
21. Key signatures and key changes are accurately performed	Disagree	Agree	Strongly Agree

Figure 3. (Continued)

22. Technical deficiency limits quality of performance	Strongly Disagree	Disagree	Agree	Strongly Agree
24. Intonation inaccuracy throughout the performance	Disagree	Agree	Strongly Agree	
26. Performance maintains good attention to dynamics	Disagree	Agree	Strongly Agree	
27. Proper balance between melody and accompaniment	Strongly Disagree	Disagree	Agree	Strongly Agree
30. Appropriate inflection at cadential points	Disagree	Agree		
31. Inconsistent connection of phrases	Disagree	Agree		
32. Articulations lack consistency in performance	Disagree	Agree	Strongly Agree	
35. Stylistic or expressive modifications (such as >, sfz, rit., ten. Cantabile) are not evident in performance	Strongly Disagree	Disagree	Agree	
36. Appropriate style is effectively communicated through performance	Disagree	Agree	Strongly Agree	
37. Some individuals disrupt the balance of the ensemble	Disagree	Agree	Strongly Agree	
39. Technical demands detract from balance of ensemble	Strongly Disagree	Disagree	Agree	

Figure 3. Final rating scale.

Tone Production			
3. *Tone Quality*	• Changes in tone quality between registers is a serious problem during performance	• Changes in tone quality between registers is a moderate problem during performance	• Changes in tone quality between registers is not a problem during performance
4. *Characteristic Balance*	• Ensemble balance between scored voices is extremely concerning during performance	• Ensemble balance between scored voices is somewhat concerning during performance	• Ensemble balance between scored voices is not concerning during performance
7. *Characteristic Tone*	• Tone is consistently undesirable due to harsh overblowing	• Tone is occasionally undesirable due to minimal harsh overblowing.	• Tone is very desirable with no harsh overblowing
8. *Air Support*	• Air support is rarely sufficient to support characteristic tone.	• Air support is sometime sufficient to support characteristic tone.	• Air support is always sufficient to support characteristic tone.
Rhythm and Pulse Accuracy			
9. *Tempo Fluctuations*	• Expressive changes in tempo and pulse are inappropriate for the style.	• Expressive changes in tempo and pulse are slightly appropriate for the style	• Expressive changes in tempo and pulse are appropriate for the style
13. *Rhythms Articulation*	• Articulations are often inconsistent with the style of music and consistently lack ensemble uniformity.	• Articulations are occasionally inconsistent with the style of music and sometimes lack ensemble uniformity.	• Articulations are consistent with style of music and maintain ensemble uniformity.
14. *Effect of Demand on Pulse*	• Technical demand often affects tempo in performance.	• Technical demand occasionally affects tempo in performance.	• Technical demand never affects tempo in performance.

Figure 4. (Continued)

15. *Rhythmic Stress and Style*	• Rhythmic stress has a major affect on the communication of proper musical style.	• Rhythmic stress has a minor affect on the communication of proper musical style.	• Rhythmic stress has no affect on the communication of proper musical style.	
16. *Steady Pulse*	• Control of pulse detracts 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.	
17. *Ensemble Releases*	• Ensemble releases are almost never executed with precision across performers.	• Ensemble releases are sometimes executed with precision across performers.	• Ensemble releases are almost always executed with precision across performers.	
18. *Rhythmic Subdivision*	• Inaccurate performance of subdivisions frequently detract from solidly communicated tempo and meter.	• Inaccurate performance of subdivisions occasionally detract from solidly communicated tempo and meter.	• Accurate performance of subdivisions contribute to solidly communicated tempo and meter.	
Pitch and Intonation Accuracy				
20. *Intonation in Expression*	• Dynamic fluctuations often detract from proper intonation.	• Dynamic fluctuations sometimes detract from proper intonation.	• Dynamic fluctuations rarely detract from proper intonation.	
21. *Key Signatures*	• Inaccurate performance of key signatures often detract from performance.	• Inaccurate performance of key signatures occasionally detract from performance.	• Accurate performance of key signature support performance.	
22. *Technical Efficiency*	• Technical efficiency is an extreme barrier to the quality of the performance.	• Technical efficiency is a moderate barrier to the quality of the performance.	• Technical efficiency somewhat of a barrier to the quality of the performance.	• Technical efficiency is not a barrier to the quality of the performance.

24. *Intonation*	• Intonation accuracy is a serious problem.	• Intonation accuracy is a moderate problem.	• Intonation accuracy is not a problem.	
Expressive Qualities / Stylistic Interpretations				
26. *Dynamics*	• Ensemble rarely demonstrates meaningful contrast in dynamics.	• Ensemble sometimes demonstrates meaningful contrast in dynamics.	• Ensemble frequently demonstrates meaningful contrast in dynamics.	
27. *Melody verses Accompaniment*	• Proper balance of the ensemble is an extreme problem in performance.	• Proper balance of the ensemble is a moderate problem in performance.	• Proper balance of the ensemble is a minor problem in performance.	• Proper balance of the ensemble not a problem in performance.
30. *Cadential Inflection*	• Phrases lack proper inflection at cadential points.		• Phrases demonstrate proper inflection at cadential points.	
31. *Connection of phrases*	• Ensemble does not meaningfully connect phrases.		• Ensemble meaningfully connects phrases.	
32. *Articulations Style*	• Articulations are often inconsistent in passages with notes of a similar style and detract much from the performance.	• Articulations are occasionally inconsistent in passages with notes of a similar style and slightly detract from the performance.	• Articulations are consistent in passages with notes of a similar style and do not detract from the performance.	
35. *Expressive Gestures*	• 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.	

Figure 4. (Continued)

36. *Musical Style*	• Appropriate musical style is rarely communicated through the ensembles performance	• Appropriate musical style is sometimes communicated through the ensembles performance	• Appropriate musical style is consistently communicated through the ensembles performance.
37. *Ensemble Balance*	• Individual ensemble members create an extreme barrier to proper ensemble balance.	• Individual ensemble members create somewhat of a barrier to proper ensemble balance.	• Individual ensemble members do not create a barrier to proper ensemble balance.
39. *Effect of Demand on Balance*	• Technical demand is an extreme barrier to proper ensemble balance.	• Technical demand is somewhat of a barrier to proper ensemble balance.	• Technical demand is not a barrier to proper ensemble balance.

Figure 4. Final rubric.

Conclusion and further research

The first research question includes the evaluation of the psychometric quality of the scale developed for the assessment of wind band performance achievement. The specific psychometric qualities that were the focus of investigation were validity, reliability, and precision of the measurement instrument. Reliability is viewed in the light of how the information gained from the data is used to distinguish between the quality and/or the amount of the latent trait, which in this case is performance achievement of wind bands. Evidence of strong reliability was seen in the high statistic for reliability of separation of items, performances, and raters (REL_{items} =.94; REL_{perf} =.98; REL_{raters} =.98). Evidence of precision is observed in the standard errors and related rating scale category diagnostics. Both the standard errors for raters and items are small, which provides strong support for high levels of precision for the scale. Furthermore, the carefully evaluated rating scale category data and related optimization provides a more precise response to the items. The scale was able to clearly distinguish between performances of varying achievement within the ensemble as displayed along an equal-interval continuum seen in the variable map. The ability of the scale to achieve this goal with reliability and precision is evidence of strong validity and assumptions for the reproducibility in further studies.

The second question guiding this research study investigated item fit to the model and variability in difficulty. These ideas specifically hone in on the content and construct validity issues facing this measurement tool. After the items were analyzed, 14 of the original 39 item stems were removed for misfit. In other words, these 14 items caused a violation of the five requirements for Rasch measurement and do not appropriately fit the model to be included in a measurement tool demonstrating invariant measurement. It is probable to consider that these 14 items demonstrate multidimensionality that confounds their effective use within the scale. Due to the items' nature of being either underfit/too predictable, or overfit/too unpredictable, they are items that would require either modification or deletion from the final scale. Because of the focus and scope of this specific study, these 14 items were deleted from the final scale. However, it is suggested that in future studies these items should be the focus of careful review such that they can be modified and transformed in an attempt to produce unidimensional items that exhibit appropriate model fit.

The third research question focused on categorical structures within each item, therefore, the remaining 25 items were reviewed for proper categorical behavior based on categorical usage, outfit, and sufficient logit separation of categories. Review of this data demonstrates strong evidence that all items in the scale did not share structural equality due to violations of monotonicity, disproportionate and skewed categorical usage, idiosyncrasies in model predictability, and poor logit separation. The resulting scale demonstrated items with categorical structures ranging from two to four categories within the Likert-type scale. This analysis provides the resulting scale with model fit that also improves precision of the scale structure thereby improving the overall construct validity of the measure.

It is important to remember the practical application of this information, or as Messick (1989) notes, the consequential validity. Consequential validity frames the implications this process has on its practical application for implementation and assessment. One of the most important initial steps needed to make this a usable assessment for large ensemble evaluations would be the development of performance standards or cut scores. These would be the defining points on the logit score of the performance achievement used to determine if an ensemble received a superior rating verses an excellent rating. In order for this to be done in an effective manner, it would require the cooperation of subject matter experts, psychometricians, and policy makers. This collection would have the task of devising the appropriate benchmarks for

achievement that define the achievement levels needed for effective and practical use as an assessment. It would be unwise for a reader of this study to consider using the sum scores of this evaluation in place of the careful work of the collection of professionals listed above. It has been shown in the study that the items themselves are not equal in category, as the ordinal interpretation would imply. Rather, the items vary in difficulty across categories and items. Summative scoring based on the final scale produced in this study would negate the validity of the scale.

Further research in the area of item stem development would also be necessary for filling in some of the gaps in logit locations of the item stems. In viewing the variable map (Figure 2), the item column reveals several gaps in placement along the continuum. For example, a gap can be seen between the item row that contains item 2 and the item row that contains item 18. This indicates areas on the logit continuum where there is less ability to distinguish between performances. Further testing of new item stems that could be used to help fill in these gaps, as well as rewritten item stems that were removed from the final scale, may help complete a series of items that filled in the scale with fewer gaps. This would continue to strengthen the ability of our measure to distinguish between performances at all levels, thus providing stronger validity and precision of this measure.

Finally, it is important for teachers viewing this study to glean some practical understanding of performance evaluation from the given results. To do this, an investigation of which items were rated most difficult and easiest to endorse is needed. The easiest item to endorse in the final scale is "*key signatures and key changes are performed correctly.*" This may indicate that during these large ensemble evaluations proper note performance is considered to be one of the most basic skills evaluated. If a teacher is striving in their classroom to only be able to teach the pitches and rhythms then they are missing out on the advanced concepts of music. However, the hardest to endorse items, 24 and 37, describe ensemble intonation and balance. Teachers who maintain an effective classroom focus on these topics are likely to be those teachers who are leading performances that achieve at the highest levels. While it is true that one cannot forgo the teaching of notes and rhythms for a focus on intonation and balance, it is important to remember these ideas during repertoire selection. Repertoire selected should be appropriate so that the notes and rhythms are attainable in an amount of time that yields the teacher appropriate time to cover more difficult topics such as intonation and ensemble balance.

Funding

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

References

Abeles, H. F. (1971). *An application of the facet-factorial approach to scale construction in the development of a rating scale for clarinet music performance* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

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

Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C. C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3–35). Orlando, FL: Academic Press.

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

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

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

Bergee, M. J. (1987). *An application of the facet-factorial approach to scale construction in the development of a rating scale for euphonium and tuba music performance* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.

Berk, R. A. (2010). The secret to the "best" ratings from any evaluation scale. *Journal of Faculty Development, 24*(1), 37–39.

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

Buckley, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects*. Retrieved from https://www.nciea.org/library/survey-approaches-used-evaluate-educators-non-tested-grades-and-subjects

Colwell, R. (1970). *The evaluation of music teaching and learning*. Englewood Cliffs, NJ: Prentice-Hall.

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education, 25*(6), 100–114.

DeCamp, C. B. (1980). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school band performance. *Dissertation Abstracts International, 41*, 1462A.

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

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

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.

Franklin, J. O. (1979). *Attitudes of school administrators, band directors, and band students towards selected activities of the public school band program* (Unpublished doctoral dissertation). Northwestern State University of Louisiana, Natchitoches, LA.

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

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

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

Jones, H. (1986). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school vocal solo performance* (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.

Kan, A., & Bulut, O. (2014). Crossed random-effect modeling: Examining the effects of teacher experience and rubric use in performance assessments. *Eurasian Journal of Educational Research*, (57), 1–27.

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

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.

National Association for Music Education. (2015). Music model cornerstone assessment: Performing ensemble proficient. *(*August), 9. Retrieved from https://nafme.org/wp-content/files/2014/11/Music_MCA_Ensemble_Performing3.pdf

National Association for Music Education. (2016). Ensemble adjudication forms. Retrieved from http://www.nafme.org/my-classroom/ensemble-adjudication-forms/

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

Nichols, J. P. (2005). A factor analysis approach to the development of a rating scale for snare drum performance. *Dialogue in Instrumental Music Educaiton, 15*(1), 11.

O'Neal, C. (2012). *Data-driven decision making: A handbook for school leaders*. Eugene, OR: International Society for Technology in Education [ISTE].

Pazitka-Munroe, W. L. (2003). *The development and validation of an audition instrument to measure vocal performance of college singers auditioning for choral ensembles* (Unpublished doctoral dissertation). Indiana University, Bloomington, IN..

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

Russell, B. E. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education, 184*, 21–34.

Sherman, C. (2006). A study of current strategies and practices in the assessment of individuals in high school bands. (Ed.D., Teachers College Columbia Univ., 2006). *Dissertation Abstracts International Section A: Humanities & Social Sciences, 67*(10), 3751.

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education, 55*(3), 268–280.

Swan, G., & Mazur, J. (2011). Examining data driven decision making via formative assessment: A confluence of technology, data interpretation heuristics and curricular policy. *Contemporary Issues In Technology And Teacher Education (CITE Journal), 11*(2), 205–222.

Sweeney, C. R. (1998). A description of student and band director attitudes toward concert band competition (Unpublished master's thesis). University of Miami, Coral Gables, FL.

U.S. Department of Education. (2009, November). *Race to the Top program: Executive summary*. Washington, DC: Author.

Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson, SC: Clemson International Institute for Tourism & Research Development.

Wayman, J. C. (2005). Involving teachers in data-driven decision-making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk, 10*(3), 295–308.

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

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal, 101*, 77–85.

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, 29*(2), 147–170.

Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q. R., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., Johns, P. M., McCarver, C. J., Schaps, R. E., Sorrell, G. W., & 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.

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

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

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

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.

Author biographies

Andrew S. Edwards is a high school music technology teacher at Peachtree Ridge High School in Suwanee, Georgia.

Kinsey E. Edwards is a middle school orchestra teacher at Alton C. Crews Middle School in Lawrenceville, Georgia. Their primary research interests include scale development and assessment as it applies to large ensemble performance evaluation.

Brian C. Wesolowski is an Associate Professor of Music Education at the University of Georgia, Hugh Hodgson School of Music. His primary research interest includes the study of rater behavior, scale development, policy of educational assessment, and broad applications of assessment, measurement, and evaluation in large-scale testing and classroom contexts.