Original Research Article

National Association for Music Education

Journal of Research in Music Education
2017, Vol. 65(1) 95–119

DOI: 10.1177/0022429417694873
jrme.sagepub.com
SAGE

The Development of a Secondary-Level Solo Wind Instrument Performance Rubric Using the Multifaceted Rasch Partial Credit Measurement Model

Brian C. Wesolowski[1], Ross M. Amend[2], Thomas S. Barnstead[3], Andrew S. Edwards[4], Matthew Everhart[5], Quentin R. Goins[6], Robert J. Grogan III[7], Amanda M. Herceg[8], S. Ira Jenkins[9], Paul M. Johns[10], Christopher J. McCarver[11], Robin E. Schaps[12], Gary W. Sorrell[13], and Jonathan D. Williams[14]

Abstract

The purpose of this study was to describe the development of a valid and reliable rubric to assess secondary-level solo instrumental music performance based on principles of invariant measurement. The research questions that guided this study

[1]The University of Georgia, Athens, GA, USA
[2]South Forsyth High School, Cumming, GA, USA
[3]Fannin County High School, Blue Ridge, GA, USA
[4]Peachtree Ridge High School, Suwanee, GA, USA
[5]Elbert County Comprehensive High School, Elberton, GA, USA
[6]Stephenson High School, Stone Mountain, GA, USA
[7]Barber Middle School, Acworth, GA, USA
[8]McNabb Middle School, Mt. Sterling, KY, USA
[9]Alpharetta High School, Alpharetta, GA, USA
[10]Thomas County Central High School, Thomasville, GA, USA
[11]Russell Middle School, Winder, GA, USA
[12]North Gwinnett Middle School, Sugar Hill, GA, USA
[13]Sutton Middle School, Atlanta, GA, USA
[14]Harvest Preparatory School, Columbus, OH, USA

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

included (1) What is the psychometric quality (i.e., validity, reliability, and precision) of a scale developed to assess secondary-level solo music performance? (2) Do the proposed items fit the measurement model, and if so, how do the items vary in difficulty? and (3) How does the structure of the rating scale vary across individual items? The psychometric considerations in this study included calibrations of items, persons, raters, school level, musical instrument, and rating scale structure using the Multifaceted Rasch Partial Credit Measurement Model. A 13-member cohort of music content experts participated as raters in this study. A total of 75 video performances of secondary-level solo and ensemble performances were evaluated. The result was the development of the Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO), a 30-item rubric consisting of rating scale categories ranging from two to four performance criteria. Implications for consequential validity, rater training, standard setting, and benchmarking are discussed.

Keywords
assessment, invariant measurement, Many Facet Partial Credit Model, Rasch, rubric

In a data-driven educational environment, valid and reliable empirical evidence of student achievement and standardized assessment systems are at the forefront of policy-based decisions (U.S. Department of Education, 2009). In music, the call to move toward assessment systems that demand standardized measurement instruments with clear benchmarking/illustrative exemplar systems is evidenced nationally through the ongoing work of the National Association for Music Education's (NAfME) Core Arts Standards and Model Cornerstone Assessment committees (Shuler, Norgaard, & Blakeslee, 2014) and internationally through the Associated Board of Royal Schools of Music's (ABRSM) and Australian Music Examinations Board's (AMEB) implementations of multiple exam systems. Furthermore, music educators are now subjected to teacher and curriculum effectiveness frameworks where individual student achievement data are an important evaluative component of such evaluations (Wesolowski, 2014, 2015). For teachers to be effectively evaluated, inferences drawn from student achievement measures must be valid and reliable. The validity of measures, however, has been called into question for many performance assessment contexts in the field of music, as students are often evaluated on nonperformance criteria when performance is the central focus of their classroom experiences (Wesolowski, 2015). The reason for this is dualistic (Wesolowski, 2014). First, states and districts are using measures not intended specifically for the context of music performance assessment as a mechanism for evaluating student performance achievement. In many instances, states and districts are implementing schoolwide measures of growth using externally created measurement instruments designed for the broad application of all grades and subjects (Buckley & Marion, 2011). As an example, the state of Georgia clearly maintains that new measures will not be designed for subject-specific areas because such measures developed

across multiple courses and subject areas "are not cost-effective" (Georgia Department of Education, 2012: 98). Furthermore, some states are evaluating music teachers based on their students' standardized mathematics and English Language Arts (ELA) evaluation reports (Steele, Hamilton, & Stecher, 2010). This poses serious consequential validity flaws in the evaluative system as the central purpose of a measurement instrument is predicated on the notion that it measures a specific and unique construct (Wilson, 2005). In these instances, measures intended for evaluating student performance were not intended by the measure developer to be used for evaluating teacher effectiveness (Messick, 1989). Second, because of the field's lack of validated measurement instruments designed specifically for music performance assessment contexts, music educators often are required to develop their own measures (Buckley & Marion, 2011). Although music educators are content experts, a lack of adequate training in scale development and related psychometrics poses serious construct validity flaws in evaluation systems, particularly when the instruments are designed to provide student achievement data in lieu of data derived from standardized tests. Districts and states therefore are not obtaining an accurate representation of performance achievement of secondary-level music students.

The purpose of this study was to describe the development of a valid and reliable rubric to assess secondary-level solo music performance. The research questions that guided this study include (1) What is the psychometric quality (i.e., validity, reliability, and precision) of the scale developed to assess secondary-level solo music performance? (2) Do the proposed items fit the measurement model, and if so, how do the items vary in difficulty? and (3) How does the structure of the rating scale vary across individual items?

Phenomenography and Outcome Space in the Scale Construction Process

Wilson (2005) brings to light the method of phenomenography as an important scale construction process component. Marton (1986) defines phenomenography as "a research method for mapping the qualitatively different ways in which people experience, conceptualize, perceive, and understand various aspects of, and phenomena in, the world around them" (p. 31). In particular, Wilson argues that the inclusion of phenomenography in scale construction processes occurs through the active acknowledgment and development of an "outcome space" (p. 71). An outcome space includes all qualitative decision-making processes underpinning the development of the scale. It is where the theoretical intersects with the practical and the quantitative intersects with the qualitative. More specifically, these decision-making processes include the carefully crafted definition of the construct intended for measurement, a clear definition of the descriptive components of the item design, strategic development of the item pool, and post hoc refinement of the items.

Although these considerations are not unique to scale construction processes in music, this study offers a unique approach of strategically implementing phenomenographic processes. The outcome space was an integral part of the scale construction

process via the inclusion of music content experts throughout the entirety of its development. As van der Linden (1992) notes, the state of the art of measurement related to test construction can be improved through these processes. In particular, he urges test constructors to

> start applying models in discussion with content experts . . . [and] to start pretesting items . . . we should be looking for the intersection between what content experts agree on and what fits the [measurement] model. This may be easier in psychological measurement, where we have much more theoretical background and where we construct real psychological variables. (5:33–6:18)

As is explained in the Method section, this study incorporated the active participation of a cohort of 13 music content experts (e.g., secondary-level music educators) throughout the entirety of the scale construction process. This process provided the content experts with the opportunity to (a) be engaged with and be thoughtful of the research and pedagogical literature; (b) draw on their unique classroom and teaching experiences; (c) reflect on the practicality of the items throughout the construct development, item design, and rating scale development processes; (d) actively engage with the preliminary rating scale through the evaluation of 75 distinct musical performances; (e) reflect on the scale development and rating processes through the debriefing of quantitative analysis and related results in a didactic manner; and (f) refine the rating scale through the post hoc rubric development process. The strength of this study, as van der Linden discusses, was to bring together the music practitioner community and the music measurement community to construct a measure in a way that uses language and content true to what secondary-level music educators are actively doing in their classrooms while establishing validity, reliability, and fairness in the scale construction process.

Psychometric Considerations

In the field of music education, factor-analytic (FA) approaches to rating scale construction most often are used as a psychometric mechanism for developing measurement instruments. FA methodologies have been used to accommodate instrument-specific performances (Abeles, 1973; Bergee, 1987; Jones, 1986; Nichols, 1991; Pazitka-Munroe, 2003; Russell, 2010; Zdzinski & Barnes, 2002), instrumental jazz improvisation (Horowitz, 1994; D. T. Smith, 2009), jazz improvisation achievement (Pfenninger, 1990), and ensemble performance (Cooksey, 1977; DCamp, 1980; B. P. Smith & Barnes, 2007; Wesolowski, 2016).

Applications of FA as a means for scale construction attempt to systematically establish common items to a particular performance area. Because the purpose of this study was to develop a new scale where confidence in the functioning of proposed items was not strong (e.g., exploratory investigations of local dependence and item fit were unknown), Rasch analysis was chosen to be the appropriate and most meaningful methodology (Christensen, Engelhard, & Salzberger, 2012). FA alone is arguably an

insufficient methodology for the development of measurement apparatuses for several reasons: (a) its inability to address spurious evidence (Kreiner & Christensen, 2004); (b) factor analysis models alone do not separate respondent and item properties, resulting in factor loadings and item intercepts that are sample dependent and confounded across raters (Meredith, 1993); and (c) factor analysis methods often rely on problematic assumptions of interval scale properties of item scores (Wraugh & Chapman, 2005; Wright, 1996). Therefore, the Multifaceted Rasch Partial Credit Measurement Model (MFR-PC; Linacre, 1989) was used to evaluate the psychometric quality (i.e., validity, reliability, and precision) of the proposed scale.

The major benefit of the Rasch model is that when adequate fit to the model is observed, five requirements for invariant measurement are met. The five requirements for 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; Engelhard, 2013). When the data fit the requirements of the Rasch model, then it becomes possible to support invariant measurement that also implies rater-invariant measurement of performances. In other words, construct-irrelevant factors (i.e., individual characteristics of persons, raters, or items) do not contribute any interference between observed data and expectations of the Rasch model.

The Partial Credit version of the model (Masters, 1982) adds an additional parameter to the model that allows for the investigation of rating scale structure across individual items. This formulation of the MFR-PC model makes it possible to test the null hypothesis of equidistant rating scale categories across all items. The addition of this parameter provides construct evidence of the measure through the verification of an increasingly monotonic relationship between adjacent categories (i.e., the preservation of increasingly positive ordering that establishes an intended direction of "more achievement"), acceptable discrimination between performances, appropriate distribution of frequency use by raters (i.e., multimodal use of all available rating scale categories), and levels of acceptable randomness for the stochastic process of probabilistic modeling (i.e., acceptable levels of unsystematic variability for probabilistic processes; Linacre, 2002). The PC model is specified as follows:

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \varepsilon_l - \gamma_m - \tau_{ik}, \tag{1}$$

where $\ln[P_{nijmk}/P_{nijmk-1}]$ = the probability that Performance n rated by Rater i on Item j receives a rating in category k rather than category $k-1$, θ_n = the logit-scale location

(e.g., achievement) of Performance n, λ_i = the logit-scale location (e.g., severity) of Rater i, δ_j = the logit-scale location (e.g., achievement) of Musical Instrument j, ε_l = the logit-scale location (e.g., achievement) of School Level l, γ_m = the logit-scale location (e.g., difficulty) of Item j, and τ_{ik} = the location on the logit scale where rating scale categories k and $k - 1$ are equally probable for Rater i.

The logit scale can be conceptualized as a common "ruler" that measures the unidimensional latent construct. In this case, the unidimensional construct can be defined as secondary-level solo wind instrument performance achievement. The logit-scale locations (i.e., measures) for each element (i.e., each performance, rater, item, etc.) of each parameter (i.e., performance parameter, rater parameter, item parameter) indicated in the previous model are the specific locations on the "ruler." This allows for a meaningful comparison of each parameter and/or element in the same unit. Analysis of the rating data was conducted using the computer program FACETS (Linacre, 2014).

Method

Rater Cohort of Content Experts

A cohort of 13 content experts participated in this study over a course of 6 weeks. The cohort had an average of 8.25 (SD = 5.59) years of secondary-level instrumental teaching experience. The cohort met for 4 days per week for 6 weeks. Each session lasted approximately an hour and a half. Cohort time was dedicated to (a) gleaning evaluative content from research literature, (b) developing and categorizing items, (c) gathering performance stimuli (e.g., videos of musical performances), (d) evaluating performances, (e) participating in think-aloud evaluation protocols, and (f) defining performance criteria.

Development of Initial Item Pool

In order to make the breadth of the item collection and validation process more manageable and to improve the practicality of future classroom implementation of this measure, instrumentation chosen to be evaluated was narrowed to five beginning instruments: flute, clarinet, alto saxophone, trumpet, and trombone (Duke & Byo, 2007). In particular, as Duke and Byo (2007) note, "many [educators] . . . teach homogeneous classes; others . . . teach mixed classes with two or more instruments together. . . . Double reeds, horns, tubas, and percussion all present unique problems" (p. 9). Members of the cohort were assigned into subcohorts based on a specific instrument (flute n = 3, clarinet n = 3, saxophone n = 3, trumpet n = 2, trombone n = 2) and given the task of gathering auditory, visual, and technique-related evaluative statements from instructional literature. A total of 133 descriptive statements were gathered from instructional literature related to flute, clarinet, saxophone, trumpet, and trombone (see Appendix A in the supplemental material available in the online version of the article). The descriptive statements were sorted and organized into seven a priori dimensions by the cohort: (a) technique, (b) tone, (c) articulation,

(d) visual, (e) melody, and (f) time/rhythm. Subcohorts evaluated each of the descriptive statements gathered for each instrument and made notes related to the appropriateness of the descriptions for evaluation purposes. Additionally, each subcohort took notes of redundant items and provided suggestions for editing the descriptive statements to produce short, concise, and useful item stems appropriate for a Likert-type rating scale and applicable to all five instruments. The cohort then met to discuss concerns and suggestions related to the items and make finalized decisions regarding item construction. The cohort collectively approved a total of 47 items. Individually, each subcohort then evaluated the items for directionality, marking each item as positive, negative, or neutral. Upon completion of this task, the cohort met to confirm the evaluations. One hundred percent agreement was found in the directionality of the items, finding that 22 items were phrased negatively and 25 items phrased positively. No further items were edited or removed from the item pool. The 47 items were randomized using a random number generator and paired with a 4-point Likert-type scale (see Appendix B in the supplemental material available in the online version of the article). Response alternatives included *strongly agree, agree, disagree*, and *strongly disagree*. The specific labels were chosen because they can clearly be understood by raters, provide grounds for establishing rater self-opinion, and offer a mechanism that attracts rater responses into categories moving from "less" (e.g., strongly disagree) to "more" (e.g., strongly agree) (Wright & Douglas, 1986). A 4-point scale was chosen specifically to establish a forced-choice response set. The elimination of a neutral category provided a better measure of intensity of participants' responses (Dumas, 1999) while avoiding violations of monotonicity in the rating scale category functioning (Wright, 1977).

Performance Stimuli

Members of the subcohorts collected a total of 75 videos of secondary-level solo performances (flute $n = 15$, clarinet $n = 15$, alto saxophone $n = 15$, trumpet $n = 15$, trombone $n = 15$) from district and state solo and ensemble performances. Thirty-seven videos represented middle school performances, and 38 videos represented high school performances. Each solo performance was accompanied by a piano, and the soloists performed in a standing position. The acceptability of video and audio stimuli quality was rated and verified by the cohort using the International Telecommunication Union's ITU-T Rating Scale (Union, 2004).

Rating Sessions and Rater Assessment Network

Rating sessions were conducted for four consecutive days, at the same time and in the same room, for an hour and a half per session. Video performances were displayed on a projector with stereo sound and played repeatedly until each member of the cohort was finished responding to each item. Each rater used an individual laptop connected to an online response form (i.e., Google Docs) to submit ratings. The assessment design was a complete assessment network, consisting of a completely crossed two-facet

design where each rater ($N = 13$) provided observed scores for each assessment component (i.e., Rater × Performance; Engelhard, 1997). Upon completion of the rating sessions and full collection of the rating data, negatively worded items were reverse coded before being subjected to empirical analysis.

Results

Variable Map

The variable map is provided in Appendix C (available in the online version of the article). The map is a graphical representation of the latent construct being developed in this study (i.e., secondary-level solo music performance). Each of the parameters included in the measurement model are displayed in each of the columns. The first column shows the logit scale. The second column shows each of the performance measures. An asterisk represents each performance achievement measure. The measures ranged from 2.45 logits to −0.71 logits ($M = 1.11$, $SD = 0.68$, $N = 75$). The third column represents the severity measures of the raters. The rater measures ranged from 0.82 logits to −0.94 logits ($M = .00$, $SD = .48$, $N = 13$). Severe raters are located at the top of the column and become more lenient moving further down the column. The fourth column shows the achievement measures for each musical instrument. The measures ranged from 0.43 logits to −0.26 logits ($M = .00$, $SD = .27$, $N = 5$). Higher achieving instruments are located at the top of the column and become lower achieving moving farther down the column. The fifth column shows the achievement measures for each school level. The measures ranged from 0.16 logits to −0.16 logits ($M = .00$, $SD = .16$, $N = 2$).

Summary Statistics

Summary statistics are provided in Appendix D (available in the online version of the article). As demonstrated in the table, the analysis indicated overall significant differences between performances ($\chi^2 = 9{,}509.00$, $p < .01$), raters ($\chi^2 = 5{,}121.30$, $p < .01$), items ($\chi^2 = 11{,}071.70$, $p < .01$), musical instrument ($\chi^2 = 1{,}589.80$, $p < .01$), and school level ($\chi^2 = 469.80$, $p < .01$). Reliability of separation for performances ($Rel > .99$) can be conceptually compared to Cronbach's coefficient alpha, implying that the measurement instrument is sensitive enough to distinguish between high-achieving and low-achieving performances with strong reproducibility of the logit locations. Reliability of separation for raters ($Rel > 0.99$), items ($Rel > 0.99$), instrument ($Rel > 0.99$), and school level ($Rel > 0.99$) implies that there is enough separation to confirm the construct validity of the measurement instrument. Mean square fit statistics represent the size of the random predictability within the measurement system. More specifically, they represent hypotheses of the response patterns based on expected predictability within the model. Infit mean square fit statistics (*MSQ*) represent sensitivity of the patterns of responses to on-target observations,

otherwise known as an information weighted sum of model variance. Outfit mean *MSQ* represent the sensitivity to outlying, off-target responses and are based on the sum of squares standardized residuals. Good data fit to the model is evidenced by mean square fit values (infit mean squared error [*MSE*] and outfit *MSE*) close to the expected value of 1.00. Acceptable range for productive parameter-level *MSQ* is between 0.50 and 1.50 (Wright & Linacre, 1994).

Calibration of Student Performances

The calibration of student performances is provided in Appendix E (available in the online version of the article). Larger measures indicate higher performance achievement, and lower measures indicate lower performance achievement. Performance 60 was the highest achieving performance (2.45 logits), and Performance 18 was the lowest achieving performance (–0.71 logits). Evidence of misfit elements within facets of interest is based on infit *MSE* statistics outside of the rule-of-thumb ranges of 0.80 and 1.20 logits as indicated by Wright and Linacre (1994) and Engelhard (2009). Overfitting performances include Performances 1, 6, 7, 13, 16, 23, 24, 28, 29, 31, 33, 36, 42, 51, 62, 71, and 73. Underfitting performances included Performances 35, 41, 54, 57, and 75.

Calibration of Raters

The calibration of rater performances is provided in Appendix F (available in the online version of the article). The table specifies the objective rank ordering of raters based on levels of leniency/severity. Rater 5 was the most severe (observed average = 2.64, logit measure = 0.82), and Rater 13 was the most lenient (observed average = 3.56, logit measure = −0.94). Raters 5 and 10 demonstrated muted rating patterns as evidenced by infit *MSE* less than 0.80. Raters 9, 11, and 12 demonstrated haphazard rating patterns as evidenced by infit *MSE* greater than 1.20. Rater behavior was made anonymous to the members of the cohort during ex post facto psychometric analysis to avoid interference with future engagement of the scale construction process.

Calibration of School Level

The calibration of school levels is provided in Appendix G (available in the online version of the article). This calibration provides supporting evidence of concurrent validity as the measurement instrument is able to significantly distinguish between middle school and high school performances in an order that is logically expected. High school performances demonstrated higher performance achievement (observed average = 3.34, logit measure = 0.16), and middle school performances demonstrated lower performance achievement (observed average = 3.02, logit measure = −0.16).

Calibration of Musical Instruments

The calibration of musical instruments is provided in Appendix H (available in the online version of the article). The table specifies the objective rank ordering of musical instruments based on levels of achievement. Flute performers were found to be the highest achieving (observed average = 3.42, logit measure = 0.43). Clarinet performers were found to be the lowest achieving (observed average = 2.83, logit measure = −0.26).

Calibration of Items

The calibration of items is provided in Table 1. The calibrations represent the difficulty of each item. The larger the logit measure, the more difficult the item. The most difficult item was Item 25 (accurately adjusts for standard instrument-related discrepancies in intonation) (observed average = 2.33, logit measure = 1.33), and the easiest item was Item 44 (body is slouched) (observed average = 3.73, logit measure = −1.22). Items that demonstrated overfit included Items 20, 22, 28, 30, 34, 36, 41, 43, 44, and 45. Items that demonstrated underfit included Items 6, 7, 23, 25, 27, 29, 35, 42, 46, and 48.

Rating Scale Category Diagnostics

After eliminating the misfitting items from the item pool, rating scale category structures were examined for the remaining fit items. As suggested by Linacre (2002), steps can be taken to optimize rating scale category structures. Modification of the structure based on this empirical methodology provides more rigorous examination and precise estimation of performances, ultimately addressing and improving validity issues surrounding construct validity of the measurement instrument. Additionally, this post hoc investigation can clarify the meaning of the collected data and improve subsequent use of the scale (Bond & Fox, 2015). Table 2 contains the necessary empirical data to be considered. First, frequency counts for each of the four categories were examined. Uniformly distributed frequency counts across each of the rating scale categories is optimal for the calibration of rating scale difficulties. Each of the frequency counts is above Linacre's recommendation of 10 per category; however, Item 1 (Category 1), Item 3 (Category 1), Item 10 (Category 1), Item 20 (Categories 1, 2, and 3), Item 26 (Categories 1 and 2), Item 37 (Category 1), and Item 38 (Category 1) have observed category use with low enough frequencies to produce skewed, irregular distributions. Although the observed frequency counts met Linacre's recommendation, there was still a considerable skew in the frequency use of the categories. Linacre notes that in these instances, there should be a clear "substantive pivot-point" (p. 7), where meaning of the ratings can be dichotomized into a clear bipolar response (e.g., good/bad, agree/disagree, positive/negative). Therefore, these categories were collapsed into adjacent categories to provide a more clear, substantive dichotomization between strongly disagree and disagree. Second, outfit mean squares (*MSE*) were examined for values ≥2.0. Values greater than 2.0 indicate excessive noise in the ratings. More specifically, categories exhibiting values ≥2.0 have

Table 1. Calibration of Item Facet.

Item number	Observed average	Measure	Standard error	Infit *MSE*	Standardized infit	Outfit *MSE*	Standardized outfit
25	2.33	1.23	0.04	0.78	–5.78	0.77	–5.88
17	2.37	1.12	0.04	1.00	0.10	1.05	1.12
31	2.47	0.99	0.04	1.00	0.11	1.04	0.95
2	2.52	0.88	0.04	0.87	–3.23	0.89	–2.47
4	2.57	0.84	0.04	1.19	4.44	1.29	6.32
13	2.59	0.82	0.04	1.06	1.54	1.12	2.39
21	2.63	0.77	0.04	0.80	–5.27	0.77	–5.62
23	2.70	0.67	0.04	0.78	–5.71	0.77	–5.95
27	2.71	0.61	0.04	0.66	–9.00	0.64	–9.00
33	2.74	0.58	0.04	0.90	–2.56	0.97	–0.64
12	2.74	0.54	0.04	0.89	–2.79	0.87	–3.06
6	2.81	0.51	0.04	0.74	–4.42	0.80	–4.84
42	2.82	0.45	0.04	0.79	–5.51	0.79	–4.85
46	2.82	0.45	0.04	0.76	–6.29	0.75	–6.21
1	2.88	0.38	0.04	1.06	1.45	1.04	0.82
10	2.85	0.37	0.04	0.83	–4.18	0.80	–4.86
38	2.89	0.37	0.04	0.93	–1.79	0.94	–1.22
47	2.96	0.32	0.04	0.80	–5.07	0.79	–3.99
35	2.93	0.28	0.04	0.75	–2.78	0.88	–2.63
7	2.94	0.27	0.04	0.68	–1.49	0.93	–1.43
37	2.96	0.17	0.04	0.86	–3.49	0.81	–4.27
18	3.05	0.12	0.04	0.91	–2.01	0.90	–2.11
48	3.05	0.12	0.04	0.78	–2.01	0.90	–2.11
3	3.03	–0.02	0.05	0.87	–3.18	0.86	–3.23
9	3.16	–0.02	0.04	1.16	3.31	1.41	5.78
19	3.17	–0.07	0.04	1.14	3.01	1.56	8.33
29	3.11	–0.08	0.04	0.78	–5.52	0.71	–6.01
5	3.15	–0.09	0.04	1.01	0.34	1.10	1.82
41	3.21	–0.12	0.04	1.21	4.39	1.44	5.87
32	3.32	–0.22	0.04	1.40	7.36	1.79	8.10
26	3.25	–0.28	0.05	0.87	–3.00	0.83	–3.20
22	3.36	–0.30	0.04	1.29	5.19	1.51	5.65
40	3.27	–0.31	0.05	0.92	–1.76	0.91	–1.72
28	3.45	–0.39	0.04	1.41	6.80	2.02	7.15
30	3.42	–0.42	0.04	1.31	5.37	1.30	3.19
39	3.37	–0.45	0.05	1.19	3.57	1.18	2.45
45	3.50	–0.45	0.05	1.26	4.10	1.97	7.68
8	3.44	–0.48	0.05	1.10	1.75	1.28	3.37
14	3.50	–0.53	0.05	1.16	2.73	1.58	5.54
24	3.53	–0.71	0.05	1.16	2.65	1.70	7.11

(continued)

Table 1. (continued)

Item number	Observed average	Measure	Standard error	Infit *MSE*	Standardized infit	Outfit *MSE*	Standardized outfit
16	3.60	–0.77	0.05	1.14	2.10	1.60	5.10
34	3.71	–0.80	0.05	1.56	6.16	4.58	9.00
11	3.69	–0.94	0.05	1.12	1.66	1.82	5.24
20	3.80	–0.96	0.06	1.30	2.75	4.89	9.00
36	3.73	–0.99	0.06	1.27	1.56	1.58	3.21
15	3.64	–1.10	0.06	1.11	1.73	1.60	5.84
43	3.75	–1.15	0.06	1.23	2.75	2.29	6.36
44	3.73	–1.22	0.06	1.34	0.97	1.13	1.26
Mean	3.11	0.00	0.04	1.03	–0.20	1.32	0.90
SD	0.40	0.64	0.01	0.20	4.00	0.82	5.00

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

been used by raters in unexpected contexts. Therefore, Items 5 (Category 1), 8 (Category 1), 9 (Category 1), 11 (Category 1), 14 (Categories 1 and 2), 15 (Category 1), 16 (Category 1), 19 (Category 1), 20 (Categories 1 and 2), 24 (Category 1), 32 (Category 1), and 36 (Category 1) were collapsed into adjacent categories. Third, average observed logit measures were examined for violations of monotonicity. Monotonocity can be described as the continuous advancement of threshold calibrations (Andrich, 1996). This is a requirement for inferential interpretability of the rating scale. In instances when incrementally higher measures were not observed, violating categories were collapsed into adjacent categories. The second categories of Item 5, Item 11, Item 14, Item 15, Item 16, Item 19, Item 24, Item 32, Item 39, and Item 44 demonstrated violations of monotonicity and therefore were collapsed.

Rubric Development and Defining Performance Criteria Descriptors

Upon completion of the psychometric analyses, the cohort was debriefed on the quantitative results of the rating process. The cohort then participated in ex post facto qualitative analyses of psychometric results through the development of rating structure descriptions. The use of content experts in this phase of measure development helped provide methodological and epistemological insight into the scale development process. First, subcohorts worked independently to qualitatively define each of the remaining rating scale categories for the remaining fit items. The instructional literature used to develop the initial item pool was revisited to guide and facilitate the development of the descriptors. Upon completion of the task, the cohort met as a group and shared their interpretations of the performance criteria. Each of the subcohort's descriptors were merged into a master rubric representing the fit items and respective rating scale categories. Collaboratively, the cohort evaluated the content for redundancy, appropriateness, and clarity and narrowed down the content to the best overall

Table 2. Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Measures, and Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit *SE*			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	133 (14)	215 (23)	236 (25)	369 (39)	–0.32 (–0.36)	0.09 (0.04)	0.48 (0.57)	1.25 (1.24)	1.20	0.80	0.90	1.20
2	182 (19)	203 (32)	253 (27)	214 (22)	–0.76 (–0.74)	–0.32 (–0.27)	0.27 (0.32)	1.14 (0.99)	1.10	0.80	0.80	0.80
3	43 (5)	195 (20)	402 (42)	313 (33)	0.08 (–0.08)	0.19 (0.33)	0.83 (0.91)	1.81 (1.65)	1.30	0.70	0.80	0.80
4	132 (14)	310 (33)	345 (36)	166 (17)	–0.48 (–0.74)	–0.28 (–0.27)	0.35 (0.36)	0.91 (1.07)	1.60	0.90	0.90	1.50
5	57 (6)	164 (17)	313 (33)	419 (44)	0.37 (–0.03)	0.31[a] (0.36)	0.72 (0.89)	1.69 (1.60)	2.00[b]	1.00	0.70	0.79
6	111 (12)	205 (22)	392 (41)	245 (26)	–0.58 (–0.50)	–0.18 (–0.06)	0.49 (0.52)	1.42 (1.24)	0.90	0.60	0.80	0.90
7	95 (10)	203 (21)	316 (33)	339 (36)	–0.30 (–0.29)	0.05 (0.11)	0.65 (0.66)	1.41 (1.36)	1.10	0.80	0.80	1.00
8	41 (4)	89 (9)	232 (24)	591 (62)	0.60 (0.25)	0.65 (0.60)	1.01 (1.09)	1.79 (1.80)	2.20[b]	1.30	0.90	1.00
9	81 (8)	158 (17)	239 (25)	475 (50)	0.10 (–0.07)	0.38 (0.30)	0.76 (0.80)	1.45 (1.49)	2.40[b]	1.20	0.80	1.20
10	93 (10)	240 (25)	334 (35)	286 (30)	–0.50 (–0.37)	–0.01 (0.05)	0.57 (0.62)	1.48 (1.33)	.80	0.70	0.80	0.80
11	21 (2)	53 (6)	124 (13)	755 (79)	1.20 (.59)	1.03[a] (0.91)	1.30 (1.38)	2.08 (2.10)	5.90[b]	1.40	0.80	1.00
12	96 (10)	269 (28)	371 (39)	217 (23)	–0.45 (–0.51)	–0.15 (–0.06)	0.45 (0.54)	1.48 (1.26)	1.00	0.80	1.00	0.80
13	217 (23)	229 (24)	236 (25)	271 (28)	–0.60 (–0.67)	–0.24 (–0.23)	0.25 (0.32)	0.98 (0.97)	1.30	0.80	1.00	1.10
14	42 (4)	77 (8)	194 (20)	640 (67)	0.79 (0.29)	0.77[a] (0.62)	0.95 (1.10)	1.80 (1.81)	3.00[b]	2.00[b]	0.80	1.00
15	11 (1)	48 (5)	214 (22)	680 (71)	2.03 (0.74)	1.25[a] (1.08)	1.49 (1.57)	2.30 (2.30)	6.90[b]	1.30	0.90	1.00
16	27 (3)	62 (7)	174 (18)	690 (72)	1.12 (0.46)	0.93[a] (0.79)	1.13 (1.27)	1.98 (1.99)	4.50[b]	1.40	0.80	1.00
17	265 (28)	261 (27)	233 (24)	194 (20)	–0.88 (–0.90)	–0.44 (–0.42)	0.16 (0.17)	0.82 (0.81)	1.20	0.90	0.90	1.10
18	79 (8)	152 (16)	361 (38)	361 (38)	–0.19 (–0.20)	0.15 (0.19)	0.64 (0.74)	1.57 (1.45)	1.00	1.00	0.70	.90
19	68 (7)	151 (16)	282 (30)	452 (47)	0.41 (–0.04)	0.37[a] (0.33)	0.72 (0.85)	1.54 (1.55)	3.50[b]	1.20	0.80	1.00

(continued)

Table 2. (continued)

Item	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit *SE*			
	1	2	3	4	1	2	3	4	1	2	3	4
20	27 (3)	21 (2)	63 (7)	842 (88)	1.39 (0.56)	1.65 (0.87)	1.25[a] (1.32)	2.00 (2.04)	9.30[b]	7.70[b]	1.40	1.10
21	171 (18)	245 (26)	306 (32)	231 (24)	–0.75 (–0.67)	–0.32 (–0.22)	0.34 (0.36)	1.23 (1.04)	.90	0.50	0.80	0.80
24	26 (3)	74 (8)	224 (24)	629 (66)	1.39 (0.43)	0.91[a] (0.77)	1.10 (1.26)	1.98 (1.98)	5.10[b]	1.70	0.70	1.00
26	42 (4)	143 (15)	306 (32)	462 (48)	0.22 (0.12)	0.37 (0.49)	0.89 (1.02)	1.85 (1.73)	1.20	0.70	0.70	0.80
31	202 (21)	296 (31)	264 (28)	191 (20)	–0.73 (–0.82)	–0.40 (–0.34)	0.18 (0.26)	1.02 (0.92)	1.30	0.70	1.00	1.00
32	68 (7)	128 (13)	190 (20)	567 (59)	0.78 (0.07)	0.59[a] (0.42)	0.67 (0.91)	1.55 (1.59)	3.90[b]	1.30	1.00	1.10
33	146 (15)	252 (26)	262 (27)	293 (31)	–0.48 (–0.51)	–0.15 (–0.08)	0.38 (0.48)	1.28 (1.15)	1.50	0.70	0.70	0.80
37	70 (7)	226 (24)	329 (35)	328 (34)	–0.15 (–0.22)	0.05 (0.20)	0.70 (0.76)	1.61 (1.47)	1.10	0.60	0.70	0.80
38	106 (11)	197 (21)	348 (37)	302 (32)	–0.36 (–0.38)	0.02 (0.03)	0.47 (0.59)	1.44 (1.30)	1.30	0.80	0.70	0.90
39	37 (4)	119 (16)	251 (26)	546 (57)	0.67 (0.24)	0.64[a] (0.60)	1.12 (1.10)	1.77 (1.81)	1.70	1.10	0.90	1.10
40	39 (4)	111 (12)	356 (37)	447 (47)	0.21 (0.13)	0.47 (0.50)	0.90 (1.03)	1.87 (1.76)	1.10	1.00	0.70	0.90
44	10 (1)	23 (2)	179 (19)	741 (78)	1.40 (0.81)	1.28[a] (1.14)	1.56 (1.63)	2.37 (2.37)	3.10[b]	1.00	0.80	1.00
47	136 (14)	169 (18)	245 (58)	403 (42)	–0.39 (–0.32)	0.00 (0.07)	0.42 (0.58)	1.40 (1.25)	1.10	0.70	0.50	0.80

Note: Category 1 = *strongly disagree*, Category 2 = *disagree*, Category 3 = *agree*, Category 4 = *strongly agree* after reverse coding.
[a]Indicates violation of monotonicity.
[b]Indicates outfit *MSE* ≥2.00.

representation. In order to convert the items of the rating scale to learning outcomes in the rubric, all directionality within the item was removed in order to establish a nondirectional, content-based learning outcome. As an example, Item 33 (extraneous body motion) was adapted to *body motion*. In the process of developing meaningful descriptors for each criterion performance level (i.e., assigning narrative to each remaining rating scale categories), it was important to map the directionality appropriately. Therefore, the structure stemming from the reverse-coded items was used. This allows for each learning outcome and related proficiency levels to be coded uniformly in the same direction while maintaining the integrity of the collapsed rating scale category structures from the rating scale.

To focus on the continuity of narrative across the criterion performance level descriptors, the cohort collectively drew their attention to a series of pre-established anchors (Vagias, 2006). The application of specific anchors to each item across the criterion performance level descriptors were voted on and adapted to the rating scale structure by majority vote. Anchor selection included the categories of problem (Items 2, 3, 4, and 44), frequency (Items 2, 3, 5, 8, 16, 31, 37, 39), appropriateness (Items 9, 15, 17, 20, 21, 24, 26, 44, 47), barrier (Item 3), and detract (Item 18, 19, 26, 33, 40). A finalized version of the Music Performance Rubric for Secondary-Level Solos (MPR-2L-SOLO) is shown in Figure 1. Degrees of proficiency labels were not used to define the columns of the rubric for two reasons. First, degrees of proficiency range from two to four categories depending on the learning outcome. Second, labeling each column is counterproductive as it implies a specific standard that has not yet been determined. This is elaborated on in the following section.

Conclusion and Future Research

The first research question referred to the psychometric quality of the scale developed to assess secondary-level solo music performance. In particular, psychometric quality represented the reliability, precision, and validity of the measure. The scale displayed overall good psychometric qualities. First, reliability of the measures was addressed by the information functions of the parameters used to estimate the value of the latent trait. High reliability of separation for persons and items served as reliability evidence. Small standard errors associated with each person and item served as strong precision evidence. The combined reliability and precision evidence indicates that the measure was able to strongly separate performances based on an ordered hierarchy along the latent continuum using an equal-interval, logit gradation. Additionally, this ordering is an important component of predictive validity as the difficulty ordering of the items can now be predicted prior to data collection with sample independence.

The second and third research questions address content and construct validity of the measure. The cohort of 13 content experts addressed content validity through the initial screening of the items and collection of information from pedagogical literature. It was verified that the items represented the intended construct of measurement and any extraneous material had been omitted. After psychometric analysis, 17 of the 47 items were found to not demonstrate acceptable levels of model fit based on the five

Technique				
3. *Finger/slide dexterity*	• Agility poses an extreme barrier to efficiency in fingers/slide when changing notes. • Fingers consistently exhibit tension in lyrical and/or technical passages. • Note changes are a serious problem in performance.	• Agility poses a moderate barrier to efficiency in fingers/slide when changing notes. • Fingers often exhibit some tension in lyrical and/or technical passages. • Note changes are a moderate problem in performance.		• Agility poses no barrier to efficiency in the fingers/slide when changing notes. • Fingers rarely exhibit tension during lyrical or technical passages. • Note changes are not a problem in performance.
5. *Coordination between tongue and fingers/slide*	• Tongue and fingers/slide are rarely coordinated during performance. • Gaps between finger/slide changes and articulation initiation consistently detract from the performance.	• Tongue and fingers/slide are sometimes uncoordinated. • Gaps between finger/slide changes and articulation initiation consistently often detract from the performance.		• Tongue and fingers/slide are consistently coordinated. • Gaps between finger/slide changes and articulation initiation consistently rarely detract from the performance.
Tone				
2. *Tone quality in varying registers*	• The quality of sound is rarely characteristic when exchanges occur between registers. • Changes in tone quality between registers are a serious problem during performance.	• The quality of sound is sometimes characteristic when exchanges occur between registers. • Changes in tone quality are a moderate problem during performance.	• The quality of sound is often characteristic when exchanges occur between registers. • Changes in tone quality are a minor problem during performance.	• The quality of sound is consistently characteristic when exchanges occur between registers. • Changes in tone quality are not a problem during performance.
16. *Tone while executing expressive gestures*	• Command of characteristic tone is consistently compromised while executing expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.	• Command of characteristic tone is sometimes compromised through various combinations of expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.		• Command of characteristic tone is rarely compromised through various combinations of expressive gestures, including but not limited to vibrato, bends, turns, trills, and tremolos.

Articulation				
38. *Consistency of articulation*	• Articulations are often inconsistent in passages with notes of a similar style and detract much from the performance. • It is typical for notes to be unnecessarily accented or stressed by inconsistent attacks.	• Articulations are sometimes inconsistent in passages with notes of a similar style but detract very little from the performance. • A few notes are unnecessarily accented or stressed by inconsistent attacks.		• Articulations are rarely inconsistent in passages with notes of a similar style and do not detract from the performance. • Notes are not accented or stressed by inconsistent attacks.
Intonation				
4. *Intonation accuracy*	• Intonation accuracy is a serious problem.	• Intonation accuracy is a moderate problem.	• Intonation accuracy is a minor problem.	• Intonation accuracy is not a problem.
Visual				
31. *Posture tension*	• Tension is consistently present.	• Tension is often present.	• Tension is sometimes present.	• Tension is rarely present.
33. *Body motion*	• Body motion detracts very much from the musical performance.	• Body motion detracts some from the musical performance.	• Body motion detracts very little from the musical performance.	• Body motion does not detract from the musical performance.
24. *Instrument angle*	• Instrument angle is inappropriate for instrument.		• Instrument angle is appropriate for instrument.	
14. *Head position*	• Head alignment is inappropriate for instrument.		• Head alignment is appropriate for instrument.	
15. *Arm position*	• Arm position is inappropriate for instrument.		• Arm position is appropriate for instrument.	
13. *Wrist position*	• Wrist position is inappropriate for instrument.		• Wrist position is appropriate for instrument.	
9. *Hand position*	• Hand position is inappropriate for instrument.		• Hand position is appropriate for instrument.	
26. *Embouchure/ flexibility*	• Embouchure is inappropriate for instrument. • Embouchure detracts from musical performance.		• Embouchure is appropriate for instrument. • Embouchure does not detract from musical performance.	
19. *Cheeks*	• Cheeks are puffy and detract from embouchure support and airflow.		• Cheeks are not puffy and do not detract from embouchure support and air flow.	

8. *Jaw movement*	• Extraneous jaw motion is consistently present in articulation.	• Extraneous jaw motion is sometimes present in articulation.	• Extraneous jaw motion is rarely present or not present in articulation.
Air Support			
37. *Breath intake*	• Breath intake is rarely full, deep, or initiated from the diaphragm (i.e. breathing through the nose, breathing through the instrument, shallow breathing initiated from the chest/shoulders).	• Breath intake is often full, deep, and initiated from the diaphragm (i.e. some breathing through the nose, some breathing through the instrument, some shallow breathing initiated from the chest/shoulders).	• Breath intake is consistently full, deep, and initiated from the diaphragm.
1. *Sufficiency of air*	• Air support often inconsistent and detracts much from the quality of the performance.	• Air support is sometimes inconsistent and detracts very little from the performance.	• Air support is rarely inconsistent and does not detract from the performance.
20. *Air support in various registers of the instrument*	• Tone is inappropriately supported at various registers of the instrument.	• Tone is appropriately supported at various registers of the instrument	
Melody			
39. *Note accuracy*	• Student often demonstrates note inaccuracy.	• Student sometimes demonstrates note inaccuracy.	• Student rarely demonstrates note inaccuracy.
47. *Communication of musical phrases*	• Phrases are inappropriately contoured.	• Phrases are appropriately contoured.	
32. *Connection of phrases*	• Does not meaningfully connect phrases.	• Meaningfully connects phrases.	
10. *Inflection at cadence points*	• Melodic line rarely demonstrates inflection at cadence points.	• Melodic line sometimes demonstrates inflection at cadence points.	• Melodic line consistently demonstrates inflection at cadence points.

Expressive Devices				
11. *Stylistically-related dynamics*	• Dynamics are rarely appropriate for the style of music being performed.	• Dynamics are sometimes appropriate for style of music being performed.	• Dynamics are often appropriate for the style of music being performed.	• Dynamics are consistently appropriate for the style of music being performed.
12. *Contrast in dynamics*	• Rarely demonstrates meaningful contrast in dynamics.	• Student sometimes demonstrates meaningful contrast in dynamics.	• Student frequently demonstrates meaningful contrast in dynamics.	
18. *Subdivision of the rhythm*	• Inaccurate performance of subdivisions detracts from solidly communicated tempo and meter.		• Accurate performance of subdivisions contributes to solidly communicated tempo and meter.	
21. *Appropriateness of tempo*	• Tempo is inappropriate throughout the performance.		• Tempo is appropriate throughout the performance.	
40. *Steadiness of pulse*	• Control of pulse detracts very much from the continuous flow of the music.	• Control of pulse sometimes detracts from the continuous flow of the music.	• Control of pulse does not detract from the continuous flow of the music.	
17. *Expressive pulse and tempo fluctuations*	• Expressive changes in tempo and pulse are inappropriate for the style and demands of the literature.	• Expressive changes in tempo and pulse are slightly inappropriate for the style and demands of the literature.	• Changes in tempo and pulse are appropriate for the style and demands of the literature.	

Figure 1. Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO).

requirements of Rasch measurement. In these instances, the items were not functioning to appropriately define the unidimensional variable. Underfitting items were found to be too predictable, and overfitting items were found to be too sporadic for the stochastic processes underscoring the measurement process. It can be hypothesized that these items represent instances of multidimensionality. As such, these items were candidates for either modification or deletion. Because modification was outside the scope of this particular study, items were deleted. Future work on developing the measure includes a principal components analysis of the residuals to examine dimensionality and make necessary modification of the items. Upon undergoing other testing conditions with revised items, reanalysis of the model may provide further insight into item functioning and further reexamination of the construct.

The third research question addressed concerns of the rating scale structure. The partial credit model as a choice of parameterization tested the null hypothesis that the set of proposed items share the same rating scale structure. The insertion of this additional parameter provided statistical and substantive evidence that all items did not share the rating scale structure equally. Clear violations of monotonicity, irregularities in observation distribution, and idiosyncrasies related to predictability were brought to light through the analysis. The modifications of the structure therefore provide a different measure with different measurement implications. Such modifications improved model fit, thereby enhancing precision of the rating scale structure and providing grounds for overall improvement of construct validity of the measure.

According to Messick (1989), only construct validity of a measure is under the constructor's control as consequential validity (i.e., value implications and social consequences) moves beyond that of the test construction stage and into the implementation stage of the assessment process. Therefore, intended consequences of the measure should be discussed. First, performance standards need to be set. In particular, the development of cut scores and achievement levels need to be defined with respect to both the content skills outlined in the rubric and the new NAfME standards. As demonstrated in Appendix E (available in the online version of the article), each evaluated performance was given a logit measure that indicates the performance's respective achievement level. This measure takes into consideration the independent variability of each of the facets included in the model: (a) performance achievement, (b) rater severity, (c) musical instrument achievement, (d) school-level achievement, (e) item difficulty, and (f) rating scale category difficulty. A limitation of the study is that these measures still lack substantive meaning. The measure is continuous in nature. Therefore, consideration of cut scores needs to be carefully considered based on the measure's intended use. As an example, if the measure is intended to separate performances that are either "passing" or "failing," "accepted" or "not accepted," "scholarship" or "no scholarship," and so on, then an agreed on logit measure needs to be realized in order to make such a dichotomous distinction. Additionally, this logit cut score may change depending on the needs and desires of the assessment context, such as (a) how many performers can "be accepted" or (b) how strict is "passing." As another example, if the measure is intended to separate performances that are "superior," "excellent," "good," "fair," or "poor" or are being used to distinguish between "full scholarship," "partial scholarship," "no

scholarship," then agreed-on logit measures need to be realized in order to make such polytomous distinctions. In all these examples, cut scores can be derived by frequency counts from the process of moving down the ordered list from most achieving to least achieving. In some instances, this may be advantageous when targeting specific numbers of performers for ensemble seat positions, numbers of scholarships, and so on. However, this may not always be advantageous for two particular instances: (a) providing a clear expectation of the quality of performance required for the assessment context and (b) using a sample-dependent approach to the assessment context. It is recommended that a cohort consisting of content experts, psychometric experts, and policy makers devise a system consisting of clear-cut scores, achievement levels, and definitions of scoring values based on specifically chosen performance exemplars and in conjunction with demands of the assessment context.

Second, because the rubric's items vary in difficulty and the rating scale category threshold and discrimination parameters vary by each individual category, it is not recommended to sum scores by items. Additionally, this practice assumes sum scores are interval-level data (i.e., measures) when they are truly ordinal-level data. Although such application may be useful in a classroom to communicate formative assessment needs to improve teaching and learning (see Appendix I, available in the online version of the article), the use of the rubric as a quantitative, summative scoring mechanism is not appropriate until performance standards are developed, tested, and validated under multiple assessment contexts. Furthermore, application of this rubric in its current form would produce a final logit score that controls for each of the parameters in the current model. An obvious limitation is the need for the raw data to be subjected to the measurement model. It is suggested that in order to apply the rubric in practical settings such as all-state ensemble auditions, solo and ensemble contest, or classroom assessment conditions, a digital online system be developed that would enable the raw data to be transformed to logit measures. The development of an online system combined with appropriately agreed-on cut scores as well as systematic rater training may provide significantly more reliable, valid, and fair assessments than the best-practice assessment scenarios currently employed.

Last, a potential source of construct-irrelevant variance in any performance assessment is rater severity. This was investigated and controlled for independently of item and person functioning through its insertion as a parameter in the model. Validity and reliability concerns stem from occasions where raters facilitate the assessment process. Variability in scoring has been shown to negatively affect the music performance assessment process due to several sources of error, including rater errors varying levels of leniency and severity (Wesolowski, Wind, & Engelhard, 2016b), varying use of rating scale categories (Wesolowski et al., 2016a), and differential rater functioning (Wesolowski et al., 2015). It is therefore important to carefully consider rater behavior and the function of evaluative cues (e.g., items and rating scale categories) set forth in assessment measures to provide a valid, reliable, precise, and fair assessment processes. Therefore, it is recommended that specialized rater training protocols be developed, tested, and implemented in conjunction with the rubric. Once the protocols are in place, the consistent management of rater behavior is necessary to provide valid, reliable, and fair assessment practices.

The MPR-2L-INSTSOLO provides important implications for the assessment of music performance. From a psychometric perspective, this is the first music performance assessment measure developed using item response theory techniques and, more specifically, Rasch measurement techniques. As such, we hope that the demonstrated methodology provides a strong foundation for the future development of much-needed measurement instruments in the field of music performance and related psychological areas of study. From a formal assessment perspective, we hope that the careful attention to concerns of validity spark both formal and informal discussions related to current assessment practices and the effects of such practices on valid, reliable, and fair measurement of student performances. Last, from a classroom perspective, we hope to see a more clear, concise, and integrated approach between authentic performance assessment and grading on students' performance achievement. More specifically, we hope this rubric establishes a sound mechanism whereby teachers can provide concise empirical data to students, parents, administrations, and policy/decision makers that clearly communicates student musical performance achievement.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The Appendices are available in the online version of this article at https://doi.org/10.1177/0022429417694873.

References

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, *21*, 246–255.

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.

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. *Dissertation Asbtracts International*, *49*(05), 1086A.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*, 3rd edition. New York, NY: Routledge.

Buckly, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects*. Retrieved from http://www.nciea.org/publication_PDFs/BuckleyMarion_ Summary of Approaches for non-tested grades.pdf

Christensen, K. B., Engelhard, G., Jr., & Salzberger, T. (2012). Ask the experts: Rasch vs. factor analysis. *Rasch Measurement Transactions*, *26*, 1373–1378.

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

DCamp, 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*. Iowa City: University of Iowa.

Duke, R. A., & Byo, J. L. (2007). *The habits of musicianship: A radical approach to beginning band*. Austin: University of Texas Center for Music Learning. Retrieved from papers3://publication/uuid/912917D1-6EF0-449F-A1B6-939466C40505

Dumas, J. (1999). *Usability testing methods: Subjective measures, part II - measuring attitudes and opinions*. Washington, DC: American Institutes for Research.

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

Engelhard, G. (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*, 585–602. doi:10.1177/0013164408323240

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

Georgia Department of Education. (2012). Overview to the 2012 TKES/LKES pilot evaluation report. Retrieved from https://www.gadoe.org/School-Improvement/Teacher andLeaderEffectiveness/Documents/ Pilot%20Report_Overview%20and%20 Report%20 Combined%201-10-13.pdf

Horowitz, R. A. (1994). *The development of a rating scale for jazz guitar improvisation performance* (Doctoral thesis) Columbia University, Teachers College, New York, NY.

Jones, H. (1986). An application of the facet-factorial appraoch to scale construction in the development of a rating scale for high school vocal solo performance. *Dissertation Asbtracts International*, *47*(4), 1230.

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics—Theory and Methods*, *33*, 1239–1276.

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*, 85–106.

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

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought*, *21*, 29–49.

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

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

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

Nichols, J. P. (1991). A factor analysis approach to the development of a rating scale for snare drum performance. *Dialogue in Instrumental Music Education*, *15*, 11–31.

Pazitka-Munroe, W. L. (2003). The development and validation of an audition instrument to measure vocal performance of college singers auditioning for choral ensembles. *Dissertation Asbtracts International*, *63*(8), 2821.

Pfenninger, R. C. (1990). The development and validation of three rating scales for the objective measurement of jazz improvisation achievement. *Dissertation Asbtracts International*, *51*(8), 2674A.

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

Shuler, S. C., Norgaard, M., & Blakeslee, M. J. (2014). The new national standards for music educators. *Music Educators Journal*, *101*, 41–49. doi:10.1177/0027432114540120

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education*, *55*, 268–280. doi:10.1177/002242940705500307

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, *57*, 217–235. doi:10.1177/0022429409343549

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). Incorporating student performance measures into teacher evaluation systems. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/issues/2010/12/pdf/student_teacher_eval.pdf

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

U.S. Department of Education. (2009). *Race to the Top program executive summary*. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson, SC: Clemson University.

van der Linden, W. J. (1992). *IRT in the 1990s—Which models work best*. Retrieved from http://www.rasch.org/audio/IRT-van-der-Linden-2.mp3

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*, 77–85. doi:10.1177/0027432114540475

Wesolowski, B. C. (2015). Tracking student achievement in music performance: Developing student learning objectives for growth model assessments. *Music Educators Journal*, *102*, 39–47. doi:10.1177/0027432115589352

Wesolowski, B. C. (2016). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music*, *44*, 324–339. doi:10.1177/0305735614567700

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*, 147–170. doi:10.1177/1029864915589014

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

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016b). Rater analyses in music performance assessment: Application of the Many Facet Rasch Model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356). Chicago, IL: Gia Publications.

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

Wraugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement*, *6*, 80–99.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, *3*(1), 3–24.

Wright, B. D., & Douglas, G. A. (1986). The rating scale model for objective measurement. Retrieved from http://www.rasch.org/memo35.pdf

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

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

Author Biographies

Brian C. Wesolowski is an assistant professor of music education at the University of Georgia, Athens, GA.

Ross M. Amend is the associate director of bands at South Forsyth High School, Cumming, GA and graduate student in music education at the University of Georgia.

Thomas S. Barnstead is director of bands at Fannin County High School, Blue Ridge, GA and graduate student in music education at the University of Georgia.

Andrew S. Edwards is director of bands at Peachtree Ridge High School, Suwanee, GA and graduate student in music education at the University of Georgia.

Matthew Everhart is director of bands at Elbert County Comprehensive High School, Elberton, GA and graduate student in music education at the University of Georgia.

Quentin R. Goins is director of bands at Stephenson High School, Stone Mountain, GA and graduate student in music education at the University of Georgia.

Robert J. Grogan III is director of bands at Barber Middle School, Acworth, GA and graduate student in music education at the University of Georgia.

Amanda M. Herceg is director of bands at McNabb Middle School, Mt. Sterling, KY and graduate student in music education at the University of Georgia.

S. Ira Jenkins is associate director of bands at Alpharetta High School, Alpharetta, GA and graduate student in music education at the University of Georgia.

Paul M. Johns is assistant director of bands at Thomas County Central High School, Thomasville, GA and graduate student in music education at the University of Georgia.

Christopher J. McCarver is band director at Russell Middle School, Winder, GA and graduate student in music education at the University of Georgia.

Robin E. Schaps is associate director of bands at North Gwinnett Middle School, Sugar Hill, GA and graduate student in music education at the University of Georgia.

Gary W. Sorrell is band director at Sutton Middle School, Atlanta, GA and graduate student in music education at the University of Georgia.

Jonathan D. Williams is director of bands at Harvest Preparatory School, Columbus, OH and graduate student in music education at the University of Georgia.

Submitted October 17, 2015; accepted June 2, 2016.