Original Research Article

National Association for Music Education

Journal of Research in Music Education
1–22

DOI: 10.1177/0022429418765482
jrme.sagepub.com
SAGE

Judgmental Standard Setting: The Development of Objective Content and Performance Standards for Secondary-Level Solo Instrumental Music Assessment

Brian C. Wesolowski[1], Myriam I. Athanas[1,2], Jovan S. Burton[1,3], Andrew S. Edwards[1,4], Kinsey E. Edwards[1,5], Quentin R. Goins[1,6], Amanda H. Irby[1], Paul M. Johns[1,7], Dorothy J. Musselwhite[1], Brian T. Parido[1,8], Gary W. Sorrell[1,9], and Jonathan E. Thompson[1,10]

Abstract

The purpose of this study was to describe the development of content and performance standards for a rubric to evaluate secondary-level solo instrumental music performance using a modified bookmark standard setting procedure. The research questions that guided this study include (1) What are the psychometric qualities of a rubric to

[1]University of Georgia, Athens, GA, USA
[2]Mabry Middle School, Marietta, GA, USA
[3]Coretta Scott King Academy, Atlanta, GA, USA
[4]Peachtree Ridge High School, Suwanee, GA, USA
[5]Alton C. Crews Middle School, Lawrenceville, GA, USA
[6]Stephenson High School, Stone Mountain, GA, USA
[7]Thomas County Central High School, Thomasville, GA, USA
[8]Clarke Middle School, Athens, GA, USA
[9]Sutton Middle School, Atlanta, GA, USA
[10]Fort Valley State University, Fort Valley, GA, USA

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

evaluate secondary-level solo instrumental music performance? (2) What is the quality of ratings obtained for the standard-setting panel of subject matter expert judges? (3) What cut scores best categorize secondary-level solo instrumental performances into four performance levels across the latent performance achievement variable? and (4) What content mastery of items best categorizes achievement in secondary-level solo music performance at each of the four performance levels? A panel of eight subject matter experts participated in the study. A 30-item rubric was used to collect the judging panel's observed responses. The collected responses were transformed to linear measures using the multifaceted Rasch partial credit model. The bookmark procedure resulted in the setting of three cut points representing minimum pass levels on a latent continuum differentiating between four performance achievement levels (rudimentary, emerging, proficient, and exemplary) with clearly defined content standards. Implications for opportunity to learn are discussed.

Keywords
assessment, invariant measurement, Rasch, rubric, standards

In today's data-driven educational climate, the need to empirically provide standards-based student achievement data that are established in a valid, reliable, and fair manner is becoming increasingly necessary, particularly in the context of music and other performing arts (State Education Agency Directors of Arts Education, 2014). Standard setting is one of the most enduring and important processes in educational assessment (Cizek, 2012b; Cizek, Bunch, & Koons, 2004; Hambelton, 2001; Hambelton & Pitoniak, 2006). The importance is predicated in large part on the consequences associated with the classification of student performances at marked performance achievement levels (Cizek, 2012a).

In many secondary-level adjudicated events in music, performance levels are often divided into five descriptive categories: (a) superior, (b) excellent, (c) good, (d) fair, and (e) poor (National Association for Music Education, 2016). The consequences of classification into these categories for both individual performance achievement (e.g., solo and ensemble evaluations) and ensemble performance achievement (e.g., district and/or state large group performance evaluations) carry great weight with students, parents, administrators, and communities (Hash, 2013). The perceived success of secondary-level music programs is often predicated on the results of these formal performance assessments (Sivill, 2004). Furthermore, perceptions of program quality and teacher effectiveness are often drawn from the classifications that individual students and/or ensembles are assigned (Boyle, 1992; Burnsed, Hinkle, & King, 1985). As an example, in the context of formal concert band performance evaluations, Kirchhoff (1988) notes,

> The band contest has become the means of assessment most often used by administrators to evaluate the effectiveness of an instrumental music program. In some school districts the rating achieved by ensembles is used by administration and by the community as a barometer of their educational success or failure. (p. 274)

The establishment of validity, reliability, and fairness in the standard-setting process therefore is of upmost importance and necessary to establish quality music performance assessment contexts.

The standard-setting process can be broadly distinguished by two parallel objectives: (a) the development of content standards and (b) the development of performance standards. Content standards can be defined as "collections of statements that describe specific desired learning outcomes" (Cizek, 2012a, p. 4). Content standards address the question, "What are students expected to know and to be able to do at specific achievement levels?" The setting of content standards includes the categorizing of qualitative descriptors of performance criteria based on each specific category of achievement. Performance standards, on the other hand, "specify what level of performance on a test is required for a test taker to be classified into a given performance category" (Cizek, 2012a, p. 4). More specifically, the process of setting performance standards, or "standard setting," can operationally be defined as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek, 1993, p. 100). Performance standards address the question, "How good is good enough?" (Livingstone & Zieky, 1982) and are specifically designed to quantitatively define the minimum passing level (MPL) of each respective achievement category.

It is important to clarify the distinction between the development of content standards as a qualitative evaluation process and the development of performance standards as a quantitative measurement process. As Engelhard and Gordon (2000) explain,

> frequently, . . . the differences between measurement and evaluation are not stressed. It is important to recognize that measurement represents the development and calibration of a set of items or tasks onto a line that represents a latent variable or construct. Evaluation, on the other hand, deals with judgments of value or worth; the standard-setting process can be viewed as a process used to make these value judgments explicit in order to set a cut score on the line that represents the construct. (p. 4)

This distinction between measurement viewed as *calibration* and standard setting viewed as *evaluation* is significant because, as demonstrated in this study, different criteria and methodologies must be used to examine the quality of these two separate processes. In other words, the standard-setting process primarily involves consideration of qualitative, evaluative criteria, only then to be followed with the support of the quantitative measurement data.

From a psychometric perspective, the development of performance standards refers specifically to the setting of empirical cut points across a continuous test score scale. As Engelhard and Gordon (2000) explained, the "line" that represents a latent construct is a theoretical representation of this continuous test score scale. According to the *Standards for Educational and Psychological Testing*, the development of these cut points "embod[ies] value judgments as well as technical and empirical considerations" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 101). Standard-setting processes that incorporate both qualitative value judgments (i.e., evaluations) and quantitative psychometric models (i.e.,

calibrations) are designed to ensure that the results are combined in a systematic, reproducible, objective, and defensible manner and certify that any decisions made or classifications assigned are based on high-quality data (Cizek, 2012a). As Cizek (2001) has observed, "standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other" (p. 5).

Hansche (1998) notes that a *comprehensive* assessment system necessitates the clear alignment of content standards and performance standards via processes that blend both quantitatively driven measurement procedures and qualitatively driven evaluation procedures. To provide a clear alignment, two important steps need to be taken. First, an abstraction of conceptualized points along the hypothetical, latent continuum needs to be determined that best provides distinctions between performance levels based solely on items and content from the perspective of a panel of subject matter experts. Second, the abstraction needs to be translated into empirical cut points via the judging panel's direct use of the measurement instrument and psychometric analysis of the panel's observed responses using a specified measurement model.

The method used in this study provides a new approach to standard setting in the context of music performance assessment that embodies the distinctions between evaluation and measurement, content standard and performance standard. The purpose of this study was to describe the development of content and performance standards for a rubric to evaluate secondary-level solo instrumental music performance using a modified bookmark standard setting procedure (BSSP; Impara & Plake, 1997). This procedure was used to set three cut points that represent MPLs on a latent continuum to differentiate between four performance achievement levels (e.g., rudimentary, emerging, proficient, exemplary) in the context of secondary-level solo instrumental music performance. A 30-item rubric (Wesolowski et al., 2017) was used to collect the judging panel's observed responses (i.e., evaluation). The collected observed responses were transformed to linear measures (i.e., calibration) using the multifaceted Rasch partial credit (MFR-PC) model. This study was guided by the following four research questions: (1) What are the psychometric qualities of a rubric to evaluate secondary-level solo instrumental music performance? (2) What is the quality of ratings obtained for the standard-setting panel of subject matter expert judges? (3) What cut scores best categorize secondary-level solo performances into four performance levels across the latent performance achievement continuum? and (4) What content mastery of items best categorizes achievement of secondary-level solo instrumental music performance at each of the four performance levels?

Judgmental Standard Setting

In formal music performance assessments such as solo and ensemble festivals, large ensemble music performance evaluations, all-state auditions, college entrance auditions, or any number of other performance-based assessments in the field of music, standards are required to provide a criterion by which judges evaluate performances. Outside of the field of music, to meet the needs of high-stakes, performance-based assessment systems, judgmental standard-setting processes have been proposed as

systematic procedures that enable the conceptualization and setting of content and performance standards (Angoff, 1971; Ebel, 1972; Jaeger, 1989, Nedelsky, 1954). Broadly, these processes consist of procedures that quantify the interaction between items contained in a measurement instrument and judgmental decisions of a panel consisting of subject matter experts (SMEs) regarding the marked achievement of various levels of performances on those items (Plake, Melican, & Mills, 1991; Shepard, 1980). More specifically, each SME is asked to qualitatively estimate the probability of a minimally competent performance for each item and quantitatively rate a minimally competent performance's marked achievement for each performance level using a specified measurement instrument. The result is a specific cut point, or a minimum pass level, for each performance level being considered.

To this point in time, current practice within the field of music education has not engaged with such rigorous standard-setting practices. Assessment in the field of music education is most often grounded in general impression marking and holistic scoring schemes (Davidson & Coimbra, 2001; Forbes, 1994; Mills, 1991). As Stanley, Brooker, and Gilbert (2002) state,

> initially, they [raters] adopt a "holistic" approach, relying on a "gut reaction," an "intuitive or emotional response which is basically one of enjoyment: Am I enjoying this playing?" This early process of global assessment frequently involves respondents arriving at a tentative grade. As one examiner noted: "I look at them and I say 'Distinction, high credit.' I have bands in my own mind and then the number is immaterial—to me the number is way more negotiable than the actual range." (p. 51)

The consequence of current practice leads to assessment contexts where, according to Davidson and Coimbra (2001), "no individual assessor is really sure of what [of] their own or another assessor's thoughts and beliefs lead to a particular decision about the performance" (p. 33). This obvious uncertainty in formal music assessment practices clearly contributes to assessment results riddled with invalidity and unfair practice.

In the context of any judgmental decision-making process, variability in judges' scores can result from two sources: (a) marked differences from experience, background, differing expertise, and vantage point (Wilson, 2005) and (b) rater error stemming from the interpretation of the items and domains within the measurement instrument itself (Engelhard, 2013). In judgmental standard-setting processes, the first source of variability is welcomed as it is more advantageous to qualitatively engage with SMEs in terms of their shared understanding of the construct being measured. The second source of variability can be evaluated using indices of model data fit (Wesolowski, Wind, & Engelhard, 2015, 2016). As described in the methodologies section, both sources of variability are addressed within the judgmental standard-setting procedure used in this study.

Bookmark Standard Setting Procedure

Item response theory (IRT) is a broad umbrella term used to describe a family of mathematical measurement models that considers observed test scores to be a function of a

latent, unobservable trait (Wesolowski, in press). The BSSP is an item-mapping procedure rooted in IRT that was developed in 1996 with the growing need to report student achievement, assessment, and accountability trends associated with the high-stakes, judgmental standards-based testing movement required by the National Association of Educational Progress (NAEP; Zieky, 2012). More notably, the BSSP became most commonly used to meet the psychometric requirements associated with No Child Left Behind (NCLB) and grew out of the need to move from a dichotomous, two-performance category system (e.g., pass/fail) to a polytomous, four-level performance category system (e.g., advanced/proficient/basic/below basic). Consequently, this paradigm shift set the foundation for the four-level performance category systems that are pervasive in today's teacher effectiveness frameworks (e.g., Danielson & McGreal, 2000; Marzano, 2007), student teacher performance assessments (e.g., edTPA, 2015), and other popular performance-based educational assessment systems. The BSSP is one of the most common and popular judgmental standard-setting procedures today in high-stakes performance assessments (Cizek, 2012a).

The BSSP is a platform for qualitatively describing the knowledge, skills, and abilities (KSAs) that should be demonstrated by performances at various performance levels while also quantitatively defining cut points on a unidimensional latent continuum that represents the specific construct being measured. It was specifically developed for educational contexts where multiple levels of mastery need to be established to demonstrate and measure student achievement, growth, and progress (Wang, 2003). The benefit of the BSSP is that it can be optimized in judgmental standard-setting contexts specifically where (a) performance-based assessments are used with constructed response testing formats such as a music performance; (b) judgmental standard-setting procedures employ a judging panel consisting of multiple SMEs with varied backgrounds and experiences; (c) performance-based measurement instruments, such as a rubric or rating scale, are used as the primary apparatus; and (d) the measurement instrument utilizes polytomous items with varied rating scale category structures across items. Mitzel, Lewis, Patz, and Green (2001) explain that "the bookmark procedure can accommodate items sampled from a domain, multiple test forms, or from a single form, provided the items can be placed on a common scale using IRT methods" (p. 253). The advantage of using the bookmark method on a single measure, such as in this study, is that the SMEs set the standard based on the actual items on which the music performances will be evaluated (Citzek et al., 2004). In the context of music performance assessment, the implementation of standards derived from a modified BSSP provides a method for setting clear standards that are measurement instrument specific and in a manner that is grounded in rigorous and appropriate psychometric practice.

Psychometric Considerations

Given the subjective nature of judges' (i.e., raters') decision-making processes in the context of assessing music performances and their specific task of participating in judgmental standard-setting procedures in this study, the evaluation of rater quality is

essential. In traditional music performance assessment scoring procedures, indices of interrater reliability, intrarater reliability, and correlation coefficients are used to assess the degree to which raters agree in overall scoring and use of the rubrics. The goal in these procedures is for the raters to use the rubrics with machinelike consistency, and any divergence in scoring is considered a source of error variance. From a more modern perspective of measurement theory, however, variability can stem from multiple sources of construct-relevant variance, such as raters' severity/leniency, for example (Engelhard, 2013; Wesolowski et al., 2016; Wilson, 2005). From this perspective, raters can be evaluated as independently acting experts, where divergence of response is expected, leading to some disagreement in their overall evaluations of the performances. As a result, rater variability is embraced as marked differences from experience, background, expertise, and perspective. In the context of judgmental standard setting that employs a panel of SMEs such as this study, it was more advantageous to qualitatively engage with the expert raters in terms of their shared understanding of the construct while quantitatively controlling for their severity/leniency.

We selected the MFR-PC (Linacre, 1989/1994) model as a suitable measurement model in this study. When specifying a rater parameter within the context of the multifaceted Rasch model (MFR), it can be characterized by five requirements of rater-invariant measurement: (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) persons and raters must be simultaneously located on a single underlying latent variable. When the data fit the requirements of the Rasch model, then it becomes possible to support rater-invariant measurement of music performances (Engelhard, 2013). The partial credit (PC) model is a special formulation of the MFR model, which extends the analysis to free the response alternatives for different rating scale categories across each item within the same measurement instrument (Wright & Masters, 1982). The rating scale structure for the measurement instrument used in this study ranged from two to four category structures (see Supplemental Figure S1 in the online version of the article). Therefore, the PC model was used to provide a separate parameterization for the rating scale structure of each item, allowing for the identification of differences in application of the rating scales across the items.

Method

Participants

Eight SMEs in the field of music teaching and learning participated in this study. At the time of the standard-setting procedures, each SME was an in-service music teacher

specializing in instrumental music education. The SMEs represented varied demographics, including teaching locales (urban, $n = 4$; suburban, $n = 3$; rural, $n = 1$), current teaching level (middle school, $n = 4$; high school, $n = 3$; collegiate, $n = 1$), years of teaching experience ($M = 7.78$, $SD = 4.21$), minimum degree level (bachelor's, $n = 4$; master's, $n = 4$), and primary instrument (woodwind, $n = 4$; brass, $n = 4$). Overall, the SMEs were favorable representatives for both the construct being measured (e.g., secondary-level music performance achievement) and the intended use of the measurement instrument (e.g., secondary-level music performance assessments). All results and decisions were made by the SMEs, not by stakeholders, administrators, higher education leaders, or others not directly involved in secondary-level instrumental music teaching and learning on a daily basis. The panel met for 4 days per week for 6 weeks. Each session lasted approximately an hour and a half. Panelists were not compensated, nor did they receive any special incentives for their participation in the study.

Measurement Instrument

The Music Performance Rubric for Secondary-Level Instrumental Solos (MPR-2L-INSTSOLO; Wesolowski et al., 2017) was used as the measurement instrument for this study. The MPR-2L-INSTSOLO is a 30-item rubric consisting of rating categories ranging from two to four performance criteria. The 30 items are embedded within eight domains: (a) technique ($n = 2$), (b) tone ($n = 2$), (c) articulation ($n = 1$), (d) intonation ($n = 1$), (e) visual ($n = 11$), (f) air support ($n = 3$), (g) melody ($n = 4$), and (h) expressive devices ($n = 6$). The measure was developed and the psychometric qualities (i.e., validity, reliability, and precision) evaluated using the MFR-PC measurement model.

Performance Stimuli

A total of 89 secondary-level video performances were collected and evaluated for this study: flute ($n = 18$), clarinet ($n = 17$), saxophone ($n = 11$), oboe ($n = 6$), bassoon ($n = 4$) trumpet ($n = 7$), trombone ($n = 12$), French horn ($n = 7$), euphonium ($n = 3$), and tuba ($n = 3$). The range in performance quality was representative of secondary-level (Grades 6–12) solo music performances. The average years of performing experience for the students participating in the study was 3.21 ($SD = 1.91$). The sample of performing students was gathered from two suburban schools (high school, $n = 1$; middle school, $n = 1$). There were no requirements for the piece of music chosen. Students were asked to perform any piece of written music they felt comfortable performing.

All music performances were recorded on a high-definition video camcorder (Sony HXR-MC2500) using a broadcast-quality condenser microphone (Rode Video Mic Pro with Rycote Lyre Shockmount). All video frames used a gray canvas backdrop in the background to remove any outside distractors from the video recording of the performance. All performances were recorded with the student in a seated position while reading their selected piece of music on a black Manhasset music stand. No student

performances were affiliated with any of the SMEs on the standard-setting panel. All students and legal guardians completed documentation of informed consent, and institutional review board approval was submitted both through the authors' university as well as the county in which the schools were located.

Standard-Setting Procedures

As described in the following, the standard-setting procedure followed in this study included four key steps: (1) operational administration of the rubric to the SMEs, (2) analysis of the quantitative data from the operational administration, (3) review of the operational administration data by the SMEs, and (4) development of performance-level descriptors (PLDs) by the SMEs. Step 4, development of PLDs by the SMEs, included four distinct processes: (a) decision on the appropriate number of performance levels, (b) development of range PLDs, (c) development of target PLDs, and (d) development of reporting PLDs.

Operational Administration

An operational administration was conducted where each of the SMEs ($N = 8$) evaluated every music performance ($N = 89$) using the MPR-2L-INSTSOLO. The rater assessment design was a complete assessment network, consisting of a completely crossed three-facet design where each rater ($N = 8$) provided observed scores for each item of the rubric for each performance (i.e., rater x item x performance) (Engelhard, 1997). The operational administration was conducted for three reasons: to (a) allow the judging panel to become familiar with the measurement instrument, (b) contextualize the standard-setting process with a wide range of performance achievement levels representative of the secondary level, and (c) collect observed scores to then calibrate each music performance's ability, each item's difficulty, and each rater's severity onto a unidimensional, interval-level continuum. The operational administration rating sessions were conducted over the course of a 5-day period. The sessions occurred at the same time and in the same room for an hour and a half per session. Video performances were projected on a large white classroom screen with stereo sound. Each performance was played three times. The range of performance times of the videos performances was from 33s to 2:35min. To electronically collect scores, the raters used individual laptops with the MPR-2L-INSTSOLO coded onto an online response form (Google Forms). The operational administration provided the opportunity to jointly calibrate the raters, performances, and items for use in the standard-setting process.

Analysis of Operational Administration Results

Following the operational administration, the observed rating data were calibrated using the MFR-PC measurement model with the FACETS computer program (Linacre, 2014). In answer to research Question 1 (What are the psychometric qualities of a rubric to evaluate secondary-level solo music performance?), evidence of the

psychometric qualities of the MPR-2L-INSTSOLO is provided through the summary statistics and analysis of fit for each of the calibrated items. In answer to Research Question 2 (What is the quality of ratings obtained for the standard-setting panel of subject matter expert judges?), evidence of rater quality is provided through the analysis of fit for each of the calibrated raters.

Presentation of Operational Administration Results to Judging Panel

The judging panel was provided with calibrations of items via a visual representation of the latent continuum consisting of the 28 fitting items from the revalidated rubric (see Supplemental Figure S2 in the online version of the article). The line included a representation of the item difficulties in a rank ordering from easiest (left) to most difficult (right). The item difficulties and calibrations were derived directly from the response probabilities of the judging panel during the operational administration. The benefit of establishing a prearranged ordering of item difficulty based on psychometric theory is that it (a) allowed the judging panel to better conceptually understand the relative difficulty of the items and related student performance achievement on the items at the various achievement levels of each student performance and (b) provided a template to work from for the development of PLDs.

Development of Performance-Level Descriptors

PLDs define the specific KSAs required of student performances to be placed into the specified categories of achievement. A frequently discussed limitation of some methodologies of PLD development is that they are often created before and not within the standard-setting process (Egan, Schneider, & Ferrara, 2012). This can result in an unintended consequence of PLDs developed in a manner that does not guide the standard-setting process or conversely, where the standard-setting process does not guide the PLD development. We used Egan et al.'s (2012) framework for the development of PLDs to directly integrate the development of PLDs into the judgmental standard-setting process. More specifically, Egan et al.'s framework is a construct-centered approach (Messick, 1994) to PLD construction that aims to develop PLDs iteratively as part of a sequence embedded within the standard-setting process. Included in the framework is the disaggregation of PLDs into four distinct yet interrelated types of PLDs integrated within the judgmental standard-setting process: (a) policy PLDs, (b) range PLDs, (c) target PLDs, and (d) reporting PLDs. Because this study was not guided by policymakers or facilitated from a policy maker/stakeholder perspective, policy PLDs were deemed not relevant to this study and were therefore not considered as part of the PLD development process.

Number of performance levels. The panel openly discussed and considered "the assessment, instructional, and reporting uses to be made" (Rabinowitz, Roeber, Schroeder, & Sheinker, 2006, p. 26) regarding the number of performance levels and the related qualitative descriptors associated with each considered category. The panel considered

National Association for Music Education's five-category assessment system (e.g., superior, excellent, good, fair, poor) as well as other nonmusic educational assessment sources (e.g., National Association for Education Progress, No Child Left Behind, College Board, Educational Testing Service, Pearson, National Assessment Governing Board, the Marzano Teacher Evaluation Model, the Danielson Teacher Evaluation Framework). A final decision was reached to use a four-performance level framework: exemplary, proficient, emerging, and rudimentary.

As a result of the decision, Jager, Hambelton, and Plake's (1995) suggested four-point scheme was considered by the judging panel for this study as a baseline to systemize the KSAs that should be considered for each of the four categories. Jager et al. recommend the following considerations for writing descriptors: (a) Level 4 (the top level) indicates general competence in all benchmarks within a standard and exceptional performance in a few, (b) Level 3 (commonly the performance standard) indicates general competence in all benchmarks within the standard, (c) Level 2 indicates general competence in most benchmarks within a standard with difficulties in some of the benchmarks, and (d) Level 1 indicates difficulties in a majority of benchmarks within a standard.

Development of range performance-level descriptors. The range PLDs describe the spectrum of achievement within each of the four categories. More specifically, range PLDs "are created by test developers to identify which aspects of items align to a particular performance level in regard to the cognitive and content rigor that has been defined" (Egan et al., 2012, p. 79). Range PLDs help provide a foundation for the cognitive theory underscoring the standard development process and help deconstruct what achievement is at the various performance levels. After evaluating the figure shown in Supplemental Figure S2 (in the online version of the article), each SME was asked to independently categorize the items based on the ranking of item difficulty and their opinion of what a performance representative of each of the categories should demonstrate marked achievement.

Development of target performance-level descriptors. Target PLDs define the expectation of a performance just entering a specific performance level. More specifically, target PLDs "are used by standard setting panelists to represent just how much a threshold or borderline student in a particular performance level should know and be able to do" (Egan et al., 2012, p. 79). Marzano and Kendall's (1996) generic performance levels for procedural benchmarks were considered as criteria for developing the target PLDs. Marzano and Kendall recommend the following considerations for writing descriptors: (a) Level 4 carries out the major processes and skills inherent in the procedure with relative ease and automaticity, (b) Level 3 carries out the major processes and skills inherent in the procedure without significant error but not necessarily at an automatic level, (c) Level 2 makes a number of errors when carrying out the processes and skills important to the procedure but still accomplishes the basic purpose of the procedure, and (d) Level 1 makes so many errors when carrying out the processes and skills important to the procedure that it fails to accomplish its purpose.

The target PLDs serve as evidence to answer Research Question 3 (What cut scores best categorize secondary-level solo performances into four performance levels across the latent performance achievement variable?).

Round 1. After evaluating the operational administration results and an initial review and discussion of the results of the range of performance descriptors provided by each SME for each of the four categories, the SMEs were introduced to the idea of a *borderline performance*. The borderline performance is one that hypothetically demonstrates the KSAs at a MPL for each category. It was made clear to the SMEs the distinction between a "cut point" and a "passing score" based on Kane's (1994) interpretation:

> It is useful to draw a distinction between the *passing score*, defined as a point on the score scale, and the *performance standard*, defined as the minimally adequate level of performance for some purpose. . . . The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version. (p. 426)

The SMEs were asked to hypothesize a borderline performance to divide each category (e.g., rudimentary/emerging, emerging/proficient, and proficient/exemplary). Then, each SME was asked to evaluate the three hypothesized borderline performances using the 28-item rubric based on their interpretation of the established categories, previous open-ended discussions, and consideration of Marzano and Kendall's (1996) descriptors.

Round 2. The SMEs were provided the data from the Round 1 borderline performance results. Specifically, an open discussion included a detailed diagnostic overview of the overall observed scores versus model expected scores (see Supplemental Table S7, Round 1, in the online version of the article), total counts and percentages of performances included in the categories using the established cut points, listening of samples of performances within the categories and areas specifically targeting around the cut point, items that performances should demonstrate mastery of for each category, pedagogical considerations, and practical considerations for the cut points and related ranges. The SMEs were instructed to break into pairs to discuss their specific satisfactions and dissatisfactions with the results. The SMEs then reconvened and discussed any further considerations. After discussion had completed, the SMEs were once again asked to evaluate the three hypothesized borderline performances using the 28-item rubric based on their interpretation of the established categories, previous open-ended discussions, consideration of Marzano and Kendall's (1996) descriptors, and results and related discussions of the results. The use of multiple rounds was to "foster convergence of views as the study progresses" (Karantonis & Sireci, 2006).

Development of reporting performance-level descriptors. The reporting PLDs are developed once the cut scores are finalized, with the intent of defining the appropriate and intended interpretation of the resulting test scores. More specifically, the reporting

PLDs "describe what students who just enter a performance level should know and be able to do consistently," where "the KSPs described in lower achievement levels are subsumed by students in the higher achievement levels" (Egan et al., 2012, p. 97). The purpose of specifying PLDs was to specifically communicate to students, administrators, parents, and other stakeholders the relationship of an empirical score to the score's KSAs. In the context of reporting PLDs, cut points were considered to be mastery thresholds:

> With mastery thresholds, one can interpret the [performance] in absolute terms referred to the content of specific items. . . . If a scale score exceeds the mastery threshold of a given item, we can infer that the [performance] has attained . . . mastery of the item. The scale in this way can be interpreted by inferring the level of skill represented by the content of items with mastery thresholds in particular regions of the scale. (Bock, Mislevy, & Woodson, 1982, p. 8)

Therefore, the reporting PLDs served as the answer to Research Question 4 (What content mastery of items best categorizes achievement of secondary-level solo music performance at each of the four performance levels?)

Results

Summary Statistics and Item Analysis From the Operational Administration (Research Question 1)

The first research question asked, "What are the psychometric qualities of a rubric to evaluate secondary-level solo instrumental music performance?" Summary statistics for the MFR-PC measurement model for performances (θ), raters (λ), items (γ), and instrument (δ) can be found in Supplemental Table S1 (in the online version of the article). The analysis indicated overall significant differences for performances, $\chi^2_{(89)} = 3{,}984.40$, $p < .01$; raters, $\chi^2_{(8)} = 1{,}563.80$, $p < .01$; items, $\chi^2_{(29)} = 6{,}151.30$, $p < .01$; and instruments, $\chi^2_{(9)} = 223.90$, $p < .01$. Expected fit statistics for facets center on 1.00 with an acceptable range of 0.04 to 1.2 for judged assessments (Linacre & Wright, 1994). Therefore, as seen in Supplemental Table S1 (in the online version of the article), evidence of overall good model data fit is provided by demonstration of reasonable item mean squared error (*MSE*) ranges for infit *MSE* and outfit *MSE* statistics. The detailed calibrations for the performance facet, item facet, and instrument facet can be found in Supplemental Tables S2 (performances), S3 (items), and S4 (instruments), respectively (in the online version of the article).

The calibration of items (Supplemental Table S3, in the online version of the article) provides an objective rank ordering of item difficulty based on a probabilistic transformation of observed scores to a linear, equal-interval log odds metric. An analysis of fit statistics indicated two misfit items: (a) Item 7, posture tension (infit *MSE* = 1.35, outfit *MSE* = 1.48) and (b) Item 8, body motion (infit *MSE* = 1.42, outfit *MSE* = 3.75). As a result, both items were removed from the remaining standard-setting process.

Table 1. Calibration of Rater Facet.

Rater Number	Observed Average	Measure	Standard Error	Infit *MSE*	Standardized Infit	Outfit *MSE*	Standardized Outfit
3	1.71	1.06	0.04	0.86	−5.52	1.00	0.01
5	1.91	0.40	0.04	1.22	7.46	1.14	3.10
4	2.00	0.07	0.04	1.00	−0.09	1.01	0.24
6	2.02	−0.01	0.04	1.08	2.73	0.98	−0.28
8	2.03	−0.05	0.04	1.02	0.80	1.21	3.80
1	2.10	−0.31	0.04	0.99	−0.18	1.11	1.87
7	2.16	−0.51	0.04	1.02	0.67	1.68	8.75
2	2.20	−0.69	0.04	0.88	−4.08	0.94	−0.86

Note: Raters are arranged in measure order from most severe to least severe. *MSE* = mean squared error.

To evaluate if the rating scale structure could be better optimized for the remaining 28 fitting items, diagnostics were conducted on the structure of the rating scale categories. Using the criteria set forth by Linacre (2002) and described in the context in the development of the MPR-2L-INSTSOLO (Wesolowski et al., 2017), we found that two changes could be made to the existing 28 items to better optimize the response categories. Item 15 (jaw movement) and Item 21 (note accuracy) both demonstrated an outfit *MSE* statistic ≥2.00 for Category 1, indicating an unacceptable level of randomness, "more misinformation than information in the observations" (Linacre, 2002, p. 9). Therefore, for both items, Category 1 was collapsed into Category 2 and rewritten to accommodate the change in rating scale structure. Supplemental Table S5 (in the online version of the article) provides the full empirical diagnostics of the rating scale structure for each item. The standard-setting procedure moved forward using the revalidated rubric with 28 items (Supplemental Figure S3 in the online version of the article).

Calibration of Raters (Research Question 2)

The second research question asked, "What is the quality of ratings obtained for the standard-setting panel of subject matter expert judges?" The calibration of raters can be found in Table 1. To provide a clear interpretation and frame of reference of the calibration of items, the rater facet was centered at 0.00 logits. The raters ranged in severity from 1.06 (Rater 3, most severe) to −0.69 (Rater 2, most lenient). Rater 5 demonstrated fit statistics of 1.22 (infit *MSE*) and 1.14 (outfit *MSE*), indicating misfit by .02 logits based on Wright and Linacre's (1994) recommended threshold of 0.08 to 1.20 for high-stakes testing conditions.

Descriptive Statistics From the Range PLD Development

The full descriptive statistics for the range PLD development is provided in Supplemental Table S6 (in the online version of the article). The SMEs demonstrated 100% consensus

Table 2. Calibration of Target Performance Level Descriptors by Round.

Round	Cut Point	Observed Average	Measure	Standard Error	Infit *MSE*	Standardized Infit	Outfit *MSE*	Standardized Outfit
Round 1	Proficient/ exemplary	2.43	3.45	.25	1.17	0.80	1.20	0.40
	Emerging/proficient	1.78	0.19	.15	0.83	−1.60	1.25	1.70
	Rudimentary/ emerging	1.41	−1.26	.16	1.15	1.40	1.04	0.30
Round 2	Proficient/ exemplary	2.26	2.19	.16	0.82	−1.80	0.64	−1.10
	Emerging/proficient	2.13	1.22	.15	0.95	−0.40	0.83	−0.60
	Rudimentary/ emerging	1.42	−1.17	.17	0.71	−3.40	0.69	−2.80

Note: *MSE* = mean squared error.

agreement on mastery of 8 items to represent the rudimentary performance-level descriptor (Items 16, 17, 13, 12, 11, 9, 14, 10), 100% consensus agreement on 11 items to represent mastery within the emerging level (Items 16, 17, 13, 12, 11, 9, 14, 10, 15, 28, 27), 100% consensus agreement on 21 items to represent mastery in the proficient level (Items 16, 17, 13, 12, 11, 9, 14, 10, 15, 28, 27, 1, 21, 2, 29, 5, 19, 22, 6, 20, 18), and 100% consensus agreement on all 28 items to represent mastery of the exemplary level.

Calibration of Target PLDs (Research Question 3)

The third research question asked, "What cut scores best categorize secondary-level solo performances into four performance levels across the latent performance achievement continuum?" The calibration of target PLDs by round is shown in Table 2. The borderline rudimentary/emerging performance was 1.42 logits for Round 1 and 1.41 logits for Round 2. The cut point for the borderline emerging/proficient performance was 2.13 logits for Round 1 and 1.78 logits for Round 2. The cut point for the borderline proficient/exemplary performance was 2.26 logits for Round 1 and 2.43 logits for Round 2. The differences between Round 1 and Round 2 are expected as much of the discussion by the SMEs was concerning the cut points between rudimentary/emerging and proficient/exemplary being too extreme and forcing performances into middle categories that may deserve to be in the outside categories, particularly the exemplary category. Evidence exists in the infit and outfit *MSE* statistics that in Round 2, the values are closer to the model expectation of 1.00, indicating more accurate scoring by the SMEs. Furthermore, smaller measures of standard error are demonstrated in Round 2, indicating more precise scoring by the SMEs. The average observed and average model expected scores for each item by cut point is shown in Supplemental Table S7 (in the online version of the article).

Development of Reporting PLDs (Research Question 4)

The fourth research question asked, "What content mastery of items best categorizes achievement of secondary-level solo instrumental music performance at each of the four performance levels?" The judging panel, upon evaluation of all the empirical evidence from the operational administration, range PLDs, target PLDs and on reflection of the qualitative discussions and considerations throughout the standard-setting process, collectively developed the reporting PLDs (see Figure 1). Music performances representative of each category were randomly selected and holistically evaluated to verify their association with the categories. Furthermore, performances located just above and below the cut points and within reason of the standard errors were additionally verified using a holistic evaluation. The eight-member panel agreed with 100% consensus that the cut points functioned from a substantive and holistic perspective.

Discussion

The purpose of this study was to describe the development of content and performance standards for a rubric to evaluate secondary-level solo instrumental music performance using a modified BSSP. The assessment of student achievement aligned with content and performance standards can provide valuable information from student-, teacher-, and accountability-centered perspectives. From a student-centered perspective, it provides a framework for understanding the requirements and/or expectations for their performance at a given timepoint. From a teacher-centered perspective, it provides important diagnostic feedback on the quality of their students' performances as well as empirical evidence of students' demonstration of growth in their application of KSAs. From an accountability-centered perspective, it provides schools, districts, states, and other stakeholders a common reference point for ensuring shared principles of learning and instruction while also offering a mechanism for comparing student and program achievement across states, districts, schools, and classrooms.

As the need to provide valid, reliable, and fair accountability evidence of student achievement and teacher effectiveness increases in music teaching and learning and other performing arts, the fields are likely to move toward assessment systems grounded in empirically based measures and standards. However, the field of music needs to be cautious of ensuring student equitability and opportunity to learn as it is imperative that the goal remains to foster successful students as opposed to imposing and reaching unrealistic standards. The process of standard setting, both broadly and specific to this study, is conducted in generalized terms, not giving consideration to diversity in appropriateness of curriculum and homogeneity of student bodies and programs. As Lehman (2014) aptly noted, "in the United States, we don't have an educational system; we have 13,809 educational systems" (p. 4). The 13,809 school districts across the United States are all diverse in their curricula, offering multiple and diverse opportunities to learn musical content across the spectrum of music students. Shuler, Brophy, Sabol, McGreevy-Nicols, and Schuttler (2016) outline important variables contributing to this diversity in students' learning opportunities across the United

Secondary Level Solo Instrumental Music Performance-Level Descriptors

Rudimentary: *A performance where fundamentals are still being developed, such as proper instrument positioning and body positioning. The performance demonstrates an execution of musical concepts at a foundational level, technical facility is in its infancy, and the performance demonstrates little to no attention to expressive devices.*

Emerging: *A performance that demonstrates developing fundamental of musical performance, but is lacking in consistency and refinement of more difficult musical fundamentals. The basic fundamentals of tone, technique, intonation, air support, and melodic structure needs more attention.*

What a minimum pass level (MPL) emerging performance will likely demonstrate that rudimentary performances will not demonstrate:

- Competency in visual domain, including proper body, instrument, and embouchure positioning;
- Developing consistency in tempo, pulse, and subdivisions of the rhythm; and
- Developing consistency in pitch clarity.

Proficient: *The performance frequently displays competency and a skill set that supports accomplishment with marked promise/potential. The performance is characterized as high level, but lacks detailed use of expressive devices.*

What a minimum pass level (MPL) proficient performance will likely demonstrate that most emerging performances will not demonstrate:

- Skillful technical facility;
- Refinement of and clarity in articulation;
- Consistency of proper intonation;
- Sufficient air support;
- Consistency of pitch clarity;
- Demonstrable awareness of phrase structure and cadence points; and
- Consistency of tempo, pulse, and subdivisions of rhythm.

Exemplary: *This performance serves as a desirable model and is characterized as the best of its type within the parameters of secondary-level instrumental performance. The performance fundamentals are highly developed and performance errors are rare. The performance demonstrates the execution of musical concepts and technical facility at an elevated level, as well as apparent attention to and appropriate execution of expressive devices such as phrasing, inflection, timing, and dynamics.*

What a minimum pass level (MPL) exemplary performance will likely demonstrate that most proficient performances will not demonstrate:

- Consistency of tone quality throughout the entire range of the instrument;
- Consistency of intonation throughout the entire range of the instrument;
- Demonstration of meaningful and stylistically appropriate expressive devices such as deviations in timing, dynamics, and inflections at cadence points; and
- Meaningful and stylistically appropriate shaping of musical phrases.

Figure 1. Reporting Performance-Level Descriptors

States, including the (a) number of minutes per week of instruction offered, (b) expertise in delivering instruction and assessing student work, (c) existence and quality of arts curriculum supported by quality resources, and (d) number of students and classes each educator is given responsibility for. The relationship between opportunity to learn and standards is therefore complicated and constrained by these variables, among many others.

A U.S. Department of Education (2010) study provided evidence that the standard of "proficient" varied from state to state in high-stakes assessment contexts, where considerable attention is given to the psychometric properties of tests and measures. More surprisingly, the category of proficient in some states was found to be equivalent to a basic/rudimentary level in other states. Current state of the art in music performance is an assessment system where standards are qualitative and based solely on the basis of judgments of professionals not trained in the specific assessment task or calibrated to the measurement instrument (as occurs in high-stakes performance-based assessments such as writing). As a result, little to no public transparency of the validity, reliability, and fairness of the assessment contexts is offered. It can be assumed, therefore, that standards vary greatly from state to state, district to district, and classroom to classroom. It is therefore critical that the field of music begin moving forward with more empirically based methodologies for standard setting across assessment contexts.

The results of this study suggest that the validation of this specific music performance assessment context using the MFR-PC model in conjunction with the BSSP provides a fruitful and promising methodology for setting content and performance standards in a meaningful way. Given the increasing need to provide evidence of student achievement in the arts, this methodology provides grounds for enhancing the validity, reliability, fairness, and overall quality of performance-based assessments in the context of music. As evidenced from the rigor demonstrated in this study, the development of standards is an in-depth and time-consuming task. However, the current data-driven educational environment demands it. Moving forward, we first recommend the development of more assessment instruments aligned with the demands of the field, beginning with the development of rubrics for the evaluation of solo string, voice, and percussion, followed by rubrics for the evaluation of the ensemble types most often evaluated formally: band, string ensemble, choral, and marching band. Once evaluation instruments are developed in a valid manner with consideration to modern psychometrics, it is then appropriate to develop standards based on the specified instrument. One important limitation is that standards vary based on students' opportunity to learn and related bureaucratic educational systems in which the students are nested. Therefore, standards, as described in this study, are not one size fits all. We recommend the processes described in this study be applied to various assessment contexts at various district, state, and/or institutional levels.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher*, *11*, 4–11. doi:10.3102/0013189X011003004

Boyle, D. J. (1992). Program evaluation for secondary school music programs. *NASSAP Bulletin*, *76*(544), 63–68. doi:10.1177/019263659207654413

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, *21*, 22–29.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*, 93–106. doi:10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2001). Conjectures on the rise and call of standards setting: An introduction to context and practice. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Erlbaum.

Cizek, G. J. (2012a). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). New York, NY: Routledge.

Cizek, G. J. (Ed.). (2012b). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practices*, *23*, 31–50. doi:10.1111/j.1745-3992.2004.tb00166.x

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Davidson, J. W., & Coimbra, D. D. C. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musicae Scientiae*, *5*, 33–53. doi:10.1177/102986490100500103

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

edTPA. (2015). *Educative assessment & meaningful support: 2014 edTPA administrative report*. Amherst, MA: Author.

Egan, K. L., Schneider, C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–53.

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

Engelhard, G., & Gordon, B. (2000). Setting and evaluating performance standards for high stakes writing assessments. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 3–14). Stamford, CT: Ablex.

Forbes, G. W. (1994). Evaluating music festivals and contests—Are they fair? *Update: Applications of Research in Music Education*, *12*, 16–20. doi:10.1177/875512339401200203

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education and Praeger Publishers.

Hansche, L. N. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I*. Washington, DC: U.S. Department of Education & The Council of Chief State School Officers.

Hash, P. M. (2013). Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review*, *114*, 163–169. doi:10.1080/10632913.2013.826035

Impara, J. C., & Plake, B. S. (1997, March). *An alternative standard setting approach: Variations on a theme by Angoff*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Washington, DC: American Council on Education.

Jager, R. M., Hambelton, R. K., & Plake, B. S. (1995). *Eliciting configural performance standards through a sequenced application of complementary methods*. Paper presented at the meetings of AERA and NCME, San Francisco.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461. doi:10.3102/00346543064003425

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25, 4–12. doi:10.1111/j.1745-3992.2006.00047.x

Kirchhoff, C. (1988). The school and college band: Wind band pedagogy in the United States. In J. T. Gates (Ed.), *Music education in the United States: Contemporary issues* (pp. 259–276). Tuscaloosa, AL: The University of Alabama Press.

Lehman, P. (2014). How are we doing? In T. S. Brophy, M. L. Lai, & H. F. Chen (Eds.), *Music assessment and global diversity: Practice, measurement, and policy* (pp. 3–17). Chicago, IL: GIA Publications, Incorporated.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press. (Original work published 1989)

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

Linacre, J. M. (2014). Facets (Version 3.71.4) [Computer software]. Chicago, IL: MESA Press.

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

Livingstone, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Services.

Marzano, R. J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., & Kendall, J. S. (1996). *A comprehensive guide to designing standards-based districts, schools, and classrooms*. Aurora, CO: Mid-Continent Regional Educational Laboratory.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23. doi:10.3102/0013189X023002013

Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, *17*, 173–181. doi:10.1080/0305569910170206

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

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

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3–19. doi:10.1177/001316445401400101

Plake, B. S., Melican, G. M., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice*, *10*, 15–16, 22, 25–26. doi:10.1111/j.1745-3992.1991.tb00188.x

Rabinowitz, S. N., Roeber, E., Schroeder, C., & Sheinker, J. (2006). *Creating aligned standards and assessment systems*. Washington, DC: CCSSO.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, *4*, 447–467. doi:10.1177/014662168000400403

Shuler, S. C., Brophy, T. S., Sabol, R., McGreevy-Nichols, S., & Schuttler, M. J. (2016). Arts assessment in the age of accountability: Challenges and opportunities in implementation, design, and measurement. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 183–216). New York, NY: Routledge.

Sivill, J. R. (2004). *Students' and directors' perceptions of high school band competitions* (Unpublished doctoral dissertation). Bowling Green State University, Bowling Green, OH.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18*, 46–56. doi:10.1177/1321103X020180010601

State Education Agency Directors of Arts Education. (2014). *National Core Arts Standards: Dance, media arts, music, theatre and visual arts*. Retrieved from http://www.nationalartsstandards.org/

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*, 231–253. doi:10.1111/j.1745-3984.2003.tb01106.x

Wesolowski, B. C. (in press). Item response theory in music testing. In T. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education*. New York, NY: Oxford University Press.

Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, . . . Williams, J. D. (2017). The development of a secondary-level solo wind instrument performance rubric using the multifaceted Rasch partial credit measurement model. *Journal of Research in Music Education*, *65*, 95–119. doi:10.1177/0022429417694873

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. (2016). Rater analyses in music performance assessment: Application of the many facet Rasch model. In T. S. Brophy, J. Marlatt, &

G. K. Ritcher (Eds.), *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356). Chicago, IL: GIA.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

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

U.S. Department of Education. (2010). *The condition of education, 2010* (NCES Report No. 2010-028, Indicator 21). Washington, DC: National Center for Education Statistics.

Zieky, M. (2012). So much has changed: A historical overview of setting cut scores. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.), London: Routledge.

Author Biographies

Brian C. Wesolowski is an assistant professor of music education at the University of Georgia, Athens. His research interests include music assessment and policy.

Myriam I. Athanas is the assistant band director at Mabry Middle School in Marietta, Georgia, and an MMEd student at the University of Georgia.

Jovan S. Burton is the director of bands at Coretta Scott King Academy in Atlanta, Georgia, and an EdD student at the University of Georgia.

Andrew S. Edwards is a music technology teacher at Peachtree Ridge High School in Suwanee, Georgia, and an EdD student at the University of Georgia.

Kinsey E. Edwards is the orchestra director at Alton C. Crews Middle School in Lawrenceville, Georgia, and an EdD student at the University of Georgia.

Quentin R. Goins is the director of bands at Stephenson High School, Stone Mountain, Georgia, and an MMEd student at the University of Georgia.

Amanda H. Irby is an MMEd student at the University of Georgia.

Paul M. Johns is the assistant director of bands at Thomas County Central High School in Thomasville, Georgia, and an MMEd student at the University of Georgia.

Dorothy J. Musselwhite is a graduate assistant and PhD candidate at the University of Georgia.

Brian T. Parido is the director of bands at Clarke Middle School, Athens, Georgia, and an MMEd student at the University of Georgia.

Gary W. Sorrell is the director of bands at Sutton Middle School in Atlanta, Georgia, and an EdD student at the University of Georgia.

Jonathan E. Thompson is the director of bands at Fort Valley State University in Fort Valley, Georgia, and an EdD student at the University of Georgia.

Submitted October 18, 2016; accepted June 30, 2017.