Original Research Article

National Association for Music Education

Journal of Research in Music Education
2018, Vol. 66(3) 338–358

DOI: 10.1177/0022429418793645
jrme.sagepub.com
SAGE

Evaluating the Psychometric Qualities of a Rating Scale to Assess Pre-Service Teachers' Lesson Plan Development in the Context of a Secondary-Level Music Performance Classroom

Dorothy J. Musselwhite[1] and Brian C. Wesolowski[1]

Abstract

The purpose of this study was to evaluate the psychometric quality (i.e., validity and reliability) of a rating scale to assess pre-service teachers' lesson plan development in the context of secondary-level music performance classrooms. The research questions that guided this study include: (1) What items demonstrate acceptable model fit for the construct of lesson plan development in the context of a secondary-level music performance classroom? (2) How does the structure of the rating scale vary across items? and (3) Does differential severity emerge for academic administrators or music education content specialists across items? Using multiple teacher effectiveness frameworks, the lesson plans in this study were evaluated using a 4-point Likert-type rating scale (e.g., *strongly agree, agree, disagree, strongly disagree*) consisting of five domains: (a) instructional planning, (b) instructional delivery, (c) differentiated instruction, (d) assessment uses, and (e) assessment strategies. Secondary-level school administrators ($n = 8$) and music education content specialists ($n = 8$) rated 32 lesson plans using a balanced incomplete assessment network. The multifaceted Rasch measurement partial credit model was used in this study. Results suggest higher rater severity among administrators than music specialists. Of the 68 potential pairwise interactions examined in the study, 5 (7.4 %) of those were found to be statistically significant, which indicates that 5 raters demonstrated differential severity

[1]The University of Georgia, Athens, GA, USA

Corresponding Author:
Dorothy J. Musselwhite, Hugh Hodgson School of Music, The University of Georgia, 250 River Road, Athens, GA 30602, USA.
Email: dmusselwhite@gmail.com

across at least one lesson plan. Implications for student teacher preparation, teacher effectiveness, and the validity of measures are discussed.

Keywords
lesson plan, rating scale, Rasch model, reliability, validity

Lesson plan development is an integral component of the teaching process (Butt, 2006; Coppola, Scricca, & Connors, 2004). In this study, lesson plan development involves defining the learning outcomes and the methodological process to reach such outcomes. The practice of preplanning objectives, assessments, appropriate materials, teaching sequences, and student pacing is important for the establishment of a learning environment conducive toward optimizing student success (Brittin, 2005; Frey, Fisher, & Moore, 2005). High-quality lesson-planning skills are associated with more successful teaching practices and higher teaching competencies (Brittin, 2005; Butt, 2006; Lane & Talbert, 2013; Miksza & Berg, 2013; Schmidt, 2005). Furthermore, a teacher's prioritized attention to planning is vital to reach the needs of diverse students (Houston & Beech, 2002). Specifically, learning to use time effectively to plan is a skill with which pre-service teachers seem to struggle (Houston & Beech, 2002).

In the context of music education, pre-service music educators often perceive the skill of learning to develop lesson plans difficult to achieve demonstrative competency (Butler, 2001; Chaffin, 2009; Conway, 2002b; Lane & Talbert, 2015; Teachout, 1997). One obstacle preventing pre-service music educators from achieving success in lesson plan development is the lack of access to or availability of clearly defined curricula aligned with national and/or state-adopted standards. As Lehman (2014) notes, "in the United States we do not have an educational system, we have 13,809 educational systems" (p. 4). Lehman's sentiment alludes to the notion that the independence exhibited at the school district level may not only influence students' varied opportunities to learn in the arts but may also affect consistency of teacher evaluations due to the lack of cross-district coherence in music curricula.

Inconsistency among districts and even within individual schools themselves often leads to a wide variety of curriculum offerings (Shuler et al., 2015). In tested subjects such as mathematics and science, clearly defined objectives, expected sequences of learning, and best practice teaching strategies are clearly defined. In these instances, curricula come from either "tried and true" best practice or research-based models implemented by the state or district (Conway, 2002a). In music, however, the sequence and strategies are often drawn from students' teaching and course experiences, first introduced at the undergraduate level and further developed as the pre-service teacher gains more professional experience through various field experiences and internships. These strategies, therefore, are refined organically through trial and error.

A central content standard of national undergraduate curricula, and more specifically undergraduate music education curricula, is for pre-service teachers to demonstrate competency designing effective musical instruction through the development of

lesson plans (Council for the Accreditation of Educator Preparation, 2016; National Association of Schools of Music, 2016). In the context of secondary-level music performance methods classes, lesson planning is often geared toward a mock student audience because many students are planning lessons for their peers (Paul, 1998). The resulting lesson plans at the pre-service level regularly reflect inconsistencies in sequencing, assessment, and methodologies of sequence-based rehearsal strategies or conceptual lessons (Lane, 2006; Schleuter, 1991; Schmidt, 2005). In addition, pre-service teachers tend to be vague in their procedural descriptions and are not specific in their learning goals for students (Brittin, 2005; Lane, 2006; Schmidt, 2005). Therefore, the evaluation and measurement of students' lesson plan writing must be integrated into the undergraduate curriculum and guided by valid and reliable measurement instruments to ensure accurate feedback relating to planning practices.

Under a strict psychometric definition, achievement in lesson plan development is not directly observable and is therefore considered to be a latent (i.e., unobservable) construct (Baghaei, 2008). Therefore, secondary observable behaviors are needed to operationally define and measure the intended construct. In the context of this study, secondary behaviors come in the form of criteria, or judgmental cues, within the measurement instrument (Wesolowski, Wind, & Engelhard, 2016). The criteria set forth within a measurement instrument operationally define the latent construct and help support construct validity arguments. To properly evaluate and measure pre-service teachers' "performances" of lesson plan development and related levels of "achievement," a validated measure is needed to outline the secondary, observable behaviors that define the construct of "lesson plan development." The purpose of this study was to evaluate the psychometric quality (i.e., validity and reliability) of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The research questions that guided this study include: (1) What items demonstrate acceptable model fit for the construct of lesson plan development in the context of a secondary-level music performance classroom? (2) How does the structure of the rating scale vary across items? and (3) Does differential severity emerge for academic administrators (e.g., principals and assistant principals) or music education content specialists (e.g., university music education faculty) across items?

Background

Teacher Accountability

The National Education Association (NEA) indicates that the implementation of high-quality teacher evaluation systems leads to better teaching practices, thereby advancing student learning (NEA, 2011). The NEA (2011) further recommends that "highly trained evaluators" should conduct the evaluation of teachers. These evaluators should use clear, rigorous standards that explicitly specify the depth of knowledge, skills, abilities, and responsibilities of teachers (NEA, 2011). Models for teacher evaluation can come from national models. such as the NEA Principles of Professional Practice, or

from state-adopted, research-based models such as the Danielson (2013), Marzano Research Laboratory (2013), Stronge (2013), and Mid-continent Research for Evaluation and Learning (McREL) (The Center for Educator Effectiveness, 2013) frameworks, for example. It is important to note that these frameworks are intended for the in-service teacher and may differ from the expectations within a pre-service teaching curriculum. However, teaching frameworks are important for the development of pre-service teachers and for expectations for achievement in the field. Specifically, for the pre-service or early-career in-service teacher, teacher effectiveness frameworks provide structure for the execution of complex tasks (Danielson, 2007).

With the implementation of new teacher certification processes such as edTPA (2015) and heavy reliance on teacher evaluation frameworks, both pre-service and early-career teachers must quickly synthesize and demonstrate marked achievement of the various framework expectations of lesson plan development. Although the edTPA does not specifically employ one of the aforementioned frameworks, pre-service teachers may benefit from an introduction to these systems. Regarding lesson plan development in particular, the overarching goal of these frameworks is to increase student achievement through the clear documentation of teaching practices and gathering evidence of student learning. Teaching, in this context, is an intricate task that links a teacher's knowledge, skills, and character to meet the educational needs of the students (The Center for Educator Effectiveness, 2013).

Lesson Planning Dimensions of Teacher Evaluation Frameworks

Lesson planning is often emphasized as a pivotal aspect of the teaching process (Akyuz, Dixon, & Stephan, 2013). Specifically, lesson planning allows for the thoughtfulness of detailed methodologies, where the teacher can continually adjust and improve instruction (Kilpatrick, Swafford, & Findell, 2001). Teacher effectiveness frameworks aim to diagnose strengths and weaknesses not only in lesson planning but also in the effectiveness of teaching practices.

There are four widely used teacher effectiveness frameworks that are pervasive in today's educational landscape: (a) Danielson's (2013) *Framework for Teaching: Evaluation Instrument*, (b) Marzano Research Laboratory's (2013) *Teacher Evaluation Model*, (c) The McREL Teacher Evaluation System (The Center for Educator Effectiveness, 2013), and (d) Stronge and Associates' (2013) *Teacher/Leader Effectiveness Performance Evaluation System* (Wesolowski, 2014). Danielson's *Framework for Teaching: Evaluation Instrument* documents aspects of teaching through data-driven analysis while concurrently promoting student learning. The first edition of the framework was published in 1996 and has since been updated to reflect the changing instructional practices and overall educational climate associated with the Common Core State Standards (U.S. Department of Education, 2009). The promotion of deep engagement and the emphasis of active learning are two key components for Danielson's *Framework*. The lesson planning dimension of the framework is organized into four domains: (a) planning and preparation, (b) the classroom environment, (c) instruction, and (d) professional responsibilities.

Marzano Research Laboratory's (2013) *Teacher Evaluation Model* is based on a number of related works on assessment stemming from educational research and theory (Marzano, 2003a, 2003b, 2006, 2007; Marzano, Frontier, & Livingston, 2011; Marzano, Pickering, & Pollock, 2001). The Marzano model has sampled thousands of students and teachers in experimental and correlational studies to determine the most effective classroom strategies as related to student achievement (Marzano Research Laboratory, 2013). Similar to the Danielson framework, the lesson plan dimension of the Marzano model is organized into four domains: (a) classroom strategies and behaviors, (b) planning and preparing, (c) reflecting on teaching, and (d) collegiality and professionalism. Each domain focuses specifically on the role of teacher effectiveness within the context of a classroom.

The McREL Teacher Evaluation System (The Center for Educator Effectiveness, 2013) is a four-component framework that focuses on evaluation and accountability to improve teacher quality. The philosophy of the system states that teacher quality is a key estimator of student success and therefore is used to decrease teacher variability and recognize ineffectiveness. Teaching is then evaluated using a scale that differentiates teacher performance and provides meaningful goals. The scale of performance ratings is similar to a 5-point Likert type scale (e.g., *developing, proficient, accomplished, distinguished*, and *not demonstrated*). The McREL system emphasizes the use of these scales or rubrics as a self-reflection tool to clearly communicate to teachers how they may improve practices to advance to the next level of proficiency. The McREL framework is also referred to by the acronym *CUES.* The *CUES* framework divides the lesson plan dimension into four components: (a) content, (b) understanding, (c) environment, and (d) support.

The Stronge and Associates (2013) *Teacher/Leader Effectiveness Performance Evaluation System* was created to address the current gap between results of evaluation and the quality of an educator's work. In addition, this system purports to combine accountability and professionalism into one process. The Stronge system has been studied through multiple experimental designs to confirm its content, construct, and criterion validity as well as its reliability (Stronge, Ward, & Xu, 2013; Virginia Department of Education, 2012). The system is intended to be customizable and adaptable, as evidenced through the varied versions of multiple state adoptions (e.g., Georgia, New Jersey, and Virginia). As an example, the state of Georgia employs 10 Performance Standards that define the dimension of lesson planning. These standards are categorized under five major domains: (a) planning, (b) instructional delivery, (c) assessment of and for learning, (d) learning environment, and (e) professionalism and communication. This system is used in a longitudinal capacity because teachers are likely to be evaluated by multiple administrators within the school building over the course of a full year. Teachers are given time to converse with administrators about components of the evaluation system that cannot be seen in the lesson plan or in the classroom on the particular day of observation. In a music performance classroom, examples of this may include professionalism, communication with parents, involvement with district or state music events, or performance of students and the program outside of daily school activities.

Although these frameworks have a practical application toward the traditional classroom, concerns of validity have been raised in the context of music teaching (Wesolowski, 2014, 2015). A disparity can potentially occur in the observation process when administrators evaluate teachers of the arts. Unless an administrator has had prior training or experience in the performing arts, a performing arts teacher may not be evaluated fairly. Music teachers may be assessed with the expectation that their classroom should mirror that of a traditional academic teacher (i.e., mathematics, science, history), for example, with more transparent differentiation. Therefore, one important research question of this study is to investigate the difference of ratings between academic administrators (i.e., principals and assistant principals) and music education content specialists (i.e., university music education faculty).

Psychometric Considerations

Item response theory (IRT) is a branch of test theory where the specific qualities of an individual or group and the qualities of specific items will have an impact on an individual's or a group's response to an item (Furr & Bacharach, 2007). The Rasch measurement model is a specific version of the one-parameter-logistic (1-PL) model under the umbrella of IRT. The Rasch measurement model was used in this study to construct a linear measure from raw scores. The benefit of Rasch measurement is that when the data adequately fit the model, invariant measurement is achieved. Engelhard and Perkins (2011) define invariant measurement through five requirements: (a) The calibration of the items must be independent of the particular persons used for calibration (i.e., person-invariant calibration of test items), (b) any person must have a better chance of success on an easy item than a more difficult item (i.e., noncrossing item response functions), (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring (i.e., item-invariant measurement of persons), (d) a more able person must always have a better chance of success on any item than a less able person (i.e., noncrossing person response functions), and (e) items must be measuring a single underlying latent variable (i.e., unidimensionality as evidenced through a variable map). These requirements are defined in the context of cognitive-based exams, where the test-taker (i.e., person) directly interacts with the items on an exam. In the context of this study, *items* refer to the rubric criteria, and *persons* refer to the lesson plans. Model-data fit is achieved when all of these requirements are met. Evidence of model-data fit is necessary for providing: (a) an interpretation of construct and content reliability of the measurement instrument (Research Question 1), (b) a definition of the locations of the thresholds for each rating scale category across each individual item (Research Question 2), and (c) evidence of systematic differential severity between rater type (e.g., academic administrators and music education content specialists) across items (Research Question 3).

The Rasch-based statistics explored in this study were calculated using *FACETS* (Linacre, 2014). Specifically, this study employs the multifaceted Rasch partial credit model (MFR-PC) (Linacre, 1989). This model requires that all achievement levels available to raters on a measurement instrument be identified and ordered prior to the

distribution of the items (Masters, 1982). These levels of achievement only indicate an ordering and do not imply any categorical weighting. The PC version of the MFR model treats each rating scale category for each item independently, providing a more precise outcome estimate than the MFR model alone. The partial credit model is as specified as follows:

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

where $\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right]$ = the natural log of the probability that Performance n rated by Rater i on Item j in level m receives a rating in category k rather than category $k-1$; θ_n = achievement level of lesson plan n; λ_i = severity of rater I; δ_j = difficulty of item j; γ_m = rater type m (e.g., academic administrator or music education content specialist); τ_{ik} = the location on the logit scale where rating scale categories k and $k-1$ are equally probable for Rater i; and $\lambda_i\gamma_m$ = interaction term between rater severity and rater type.

In this study, each of the rubric criteria contains four response levels within the rating scale structure: *strongly disagree, disagree, agree*, and *strongly agree*.

The evaluation of lesson plans is a performance-based assessment; therefore, raters are needed to mediate the assessment process. The raters in this study did not undergo any training and therefore are likely to add construct-irrelevant variability to this specific assessment context. To evaluate model data fit of the raters and control for rater variability, raters must be treated similarly in the model. Under the conditions of rater-mediated assessments, Engelhard and Perkins's (2011) requirements of invariant measurement can be extended to raters, whereby: (a) rater-invariant measurement of persons (i.e., the measurement of lesson plans must be independent of the particular raters that happen to be used for the measuring), (b) noncrossing person response functions (i.e., a higher achieving lesson plan must always have a better chance of obtaining higher ratings from raters than a less achieving lesson plan), (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular lesson plans used for calibration), (d) noncrossing rater response functions (i.e., any lesson plan must have a better chance of obtaining a higher rating from lenient raters than from more severe raters, and (e) variable map (i.e., lesson plans and raters must be simultaneously located on a single underlying latent variable) (Engelhard, 2013).

Method

Initial Item Pool Generation, Raters, and Judging Plan

Items for evaluating lesson plans were gathered from performance standards from each of the teacher evaluation frameworks (reviewed earlier, The Center for Educator Effectiveness, 2013; Danielson, 2013; Marzano Research Laboratory, 2013; Stronge & Associates, 2013; Woods, 2015). Four areas were found relevant to be assessed

using only a pre-service teacher's lesson plan of the various performance standards in each of the frameworks: (a) instructional planning, (b) instructional strategies, (c) differentiated instruction, and (d) assessment strategies. These indicators, combined with performance indicators from other frameworks, became the structure for the preliminary lesson plan rating scale (see Figure S1 in the online version of the article). Relevant items from each of the frameworks were removed and transformed into statements applicable for the assessment of a music-specific lesson plan. To inspect face validity of the criteria, the authors and one outside university music education professor screened the item pool for clarity, writing style, and redundancy. Any items that appeared unclear or redundant were removed from the overall item pool. The remaining items ($N = 34$) were listed in a randomized order.

After giving informed consent, undergraduate music education majors ($n = 32$) at a large southern university submitted anonymous lesson plans (see Figures S2, S3, S4 in the online version of this article). These students ranged from second-year students to fifth-year undergraduate students. All identifying information was removed from each lesson plan to maintain student anonymity. Lesson plans were written for both middle school and high school levels, including band, orchestra, and choral content matter. A total of 32 lesson plans were used in the study, meeting the minimum sample requirement to produce statistically stable measures with a 95% confidence interval (Linacre, 1994).

The lesson plans were sent to 16 volunteer raters: (a) university music education faculty ($n = 8$) and (b) academic administrators (principals, $n = 1$; assistant principals, $n = 7$). Raters were solicited based on reputation, record of success within their field, and availability. Accompanying each lesson plan was the initial rating scale (Figure S1 in the online version of the article). Each rater independently evaluated each of four lesson plans using the 34 rating scale items on the included rating scale. The rating scale structure for each item was based on a 4-point Likert-type scale. The response alternatives included: *strongly agree, agree, disagree*, and *strongly disagree*. A 4-point rating scale structure was chosen specifically due to its absence of a neutral category, thereby requiring a forced choice, resulting in a better estimate of raters' attitudes (Dumas, 1999; Wright, 1977).

The rating scale was entered into a Google form. All raters were given explicit instructions as to the use of the form. In addition, the authors sent copies of each numbered lesson plans to the rater before evaluation. Within the form was a statement allowing the researcher to agree to terms regarding the number of lesson plans, the content of the Google form, the collection of anonymous data, and the option to not participate in the study. Raters then selected whether they consented to take part in the study. This study was granted approval by the University of Georgia Institutional Review Board.

Rater Judging Plan

The judging plan was a balanced incomplete assessment network (Engelhard, 1997). This judging plan ensures reliability and validity both within and between facets, as

recommended by Linacre and Wright (2004) and Wright and Stone (1979). Each rater evaluated four lesson plans. For example, Rater 1 evaluated Lesson Plans 1, 2, 3, and 4. In this particular judging plan, overlap is needed to ensure there is no bias in the rating. So, Rater 2 evaluated Lesson Plans 3, 4, 5, and 6. This pattern continued until Rater 16 evaluated Lesson Plans 31, 32, 1, and 2, at which point the circuit was complete. Therefore, every lesson plan was evaluated twice by each type of rater, and no single rating weighed more heavily than another. These lesson plans have been linked sufficiently based on a sound data collection design (Engelhard, 1997; Kirk, 1995; Wind, Engelhard, & Wesolowski, 2016). Linking enables data from different students and different raters to be analyzed together, thereby constructing a single measure ("Linking," 2007). This form of an incomplete assessment network was verified to demonstrate the best model-data fit among multiple incomplete assessment network structures (Wesolowski et al., 2016).

Wright Map

The Rasch model indicates its unidimensionality by displaying all facets on a linear scale. This display is called the Wright map, which depicts the operational definition of the latent construct. The Wright Map displays lesson plan difficulty, rater severity, item difficulty, and rater type on one scale (see Appendix A in the online version of this article). The first column of the Wright map is the logit-scale measure, which is the underlying scale for all facets. This scale is composed of equally spaced units representing the unidimensional latent construct. The second column indicates the distribution of lesson plans using asterisks, from high achieving to low achieving. The third column is the location of raters, from most severe to most lenient. The fourth column is location of rater type, from most severe to most lenient. The fifth column is the location of items from the rating scale, from most difficult to easiest.

Results

In this study, the MFR-PC model was used to evaluate the validity and reliability of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The descriptions provided in this section are focused on separation, as evidenced through chi-square statistics and their related reliability of separation statistics, model-data fit, and logit-scale locations, shown on the Wright map and through the calibration of elements (i.e., each lesson plan, each rater, each item) within each facet.

Summary Statistics

Appendix B (in the online version of the article) provides the summary statistics for the MFR-PC model using *FACETS* (Linacre, 2014) for lesson plans (θ), raters (λ), items (δ), and rater type (γ). The analysis indicated overall significant differences for lesson plans ($\chi^2 = 499.2$, $p < .01$), raters ($\chi^2 = 621.8$, $p < .01$), items ($\chi^2 = 296.2$,

$p < .01$), and rater type ($\chi^2 = 118.5, p < .01$). Reliability of separation is also reported for each facet. Specifically, reliability of separation refers to the reproducibility of the relative measure location (Linacre, 2017). This characteristic is interpreted similarly to Cronbach's alpha in its estimation of the spread of elements within a facet. Overall, high reliabilities of separation between lesson plans ($REL_{\text{LessonPlans}} = .94$), raters ($REL_{\text{Raters}} = .97$), items ($REL_{\text{Items}} = .89$), and rater type ($REL_{\text{RaterType}} = .98$) indicate that the Lesson Plan Evaluation Rating Scale was able to reliably separate each facet from the underlying latent trait of lesson plan achievement. More specifically, the lesson plans were able to be reliably separated at varying achievement levels across the unidimensional continuum. Both raters (.97) and items (.89) were able to separate lesson plans based on variability in achievement with reasonable reliability. Regardless of rater type, varying achievement levels of lesson plans were able to be distinguished.

Model-Data Fit. Fit statistics indicate the degree to which invariant measurement is achieved. Specifically, infit and outfit statistics are used to determine how invariant the data are. Infit mean squares refers to the fit of the data that is sensitive to inliers, focusing on individual person responses (Linacre, 2002). Outfit mean squares look at the fit of the data in response to outliers, focusing on potential effects on response patterns. Overall, mean squares seek to determine randomness that occurs in the data set. Table 1 and Appendix B (in the online version of the article) indicate that the mean infit and outfit MSE are centered near 1.00. In the strictest view, infit and outfit statistics should fall within the range of 0.8 to 1.2, indicating invariance among the data. If a statistic falls outside of this invariant range, the statistic and its related element should be carefully evaluated (Wright & Linacre, 1994). An indication of good model-data fit is evidenced through fit statistics falling within Wright and Linacre's specified range. As a result, evidence of good model-data fit indicates a degree of reasonable invariant measurement that produces interpretable estimates of measurement. When invariant measurement is achieved, along with high reliability of separation, we can infer the trustworthiness of the score interpretation (Baghaei, 2008). More broadly, the presence of invariant measurement, and therefore the trustworthiness of score interpretation, yields a strong argument for construct validity, as depicted in the Wright map (see Appendix A in the online version of the article).

Appendix C (in the online version of the article) indicates the function of the rating scale categories for each item. In other words, this table shows how raters used the categories from each item. Items are listed in numerical order, as they appeared to each rater. Columns 2 through 5 indicate the raw score of instances when each category was used on a particular item. The percentage is shown in parentheses. Columns 6 through 9 indicate the average observed measure. This number indicates where the item falls on the logit scale. Columns 10 through 13 detail each item's outfit MSE per category. Again, this statistic should fall in the range of 0.8 to 1.2, so misfit items may be detected here. Misfit items include but are not limited to: Item 3 in Category 1, Item 5 in Category 1, and Item 8 in Category 4. Each category that is found to be misfit is first

Table 1. Calibration of Items.

Item Number	Observed Average Rating	Measure	*SE*	Infit MSE	Standardized Infit MSE	Outfit MSE	Standardized Outfit MSE
9	2.13	0.98	0.21	0.74	−1.50	0.81	−1.00
10	2.27	0.95	0.22	0.76	−1.50	0.74	−1.60
25	2.17	0.77	0.20	0.87	−0.60	0.93	−0.30
28	2.28	0.74	0.21	0.66	−2.20	0.65	−2.20
29	2.25	0.73	0.21	0.82	−1.00	0.86	−0.80
16	2.25	0.73	0.19	1.18	1.00	1.22	1.20
21	2.27	0.72	0.21	0.81	−1.10	0.81	−1.10
27	2.20	0.68	0.21	0.84	−0.90	0.84	−0.80
17	2.34	0.55	0.20	1.03	0.20	1.04	0.20
20	2.27	0.49	0.20	0.62	−2.40	0.60	−2.50
33	2.31	0.44	0.20	0.67	−2.10	0.66	−2.10
34	2.48	0.43	0.24	1.15	0.80	1.08	0.40
15	2.30	0.42	0.20	1.07	0.40	1.14	0.80
4	2.91	0.39	0.26	1.18	1.00	1.20	1.00
31	2.92	0.36	0.27	1.00	0.00	0.97	0.00
8	2.30	0.34	0.19	1.34	1.80	1.42	2.10
32	2.34	0.29	0.22	0.97	−0.10	1.01	0.10
1	2.98	0.07	0.23	1.10	0.60	1.14	0.80
22	2.47	−0.03	0.19	1.27	1.50	1.39	2.10
18	2.61	−0.07	0.21	1.63	2.90	1.56	2.50
19	2.61	−0.14	0.19	0.83	−0.90	0.87	−0.70
11	2.66	−0.16	0.20	0.89	−0.50	0.82	−0.90
2	3.06	−0.24	0.25	1.00	0.00	0.99	0.00
26	2.72	−0.28	0.23	1.03	0.10	0.98	0.00
23	3.06	−0.48	0.33	1.12	0.50	1.22	0.80
6	2.81	−0.66	0.21	0.93	−0.30	0.94	−0.20
7	2.86	−0.69	0.22	1.05	0.30	1.00	0.00
13	3.14	−0.72	0.28	1.19	1.00	1.20	0.90
24	2.81	−0.82	0.23	1.19	1.00	1.12	0.60
30	2.91	−0.92	0.26	1.10	0.50	1.05	0.20
5	2.95	−0.96	0.24	1.09	0.40	0.99	0.00
14	2.86	−1.05	0.27	1.17	0.80	1.16	0.70
12	3.20	−1.14	0.28	0.89	−0.50	0.83	−0.80
3	3.11	−1.71	0.25	1.02	0.10	1.04	0.20
M	2.61	0.00	0.23	1.01	0.00	1.01	0.00
SD	0.34	0.70	0.03	0.21	1.20	0.22	1.20

Note. Items are ordered according to measure, from highest achieving to lowest achieving.

evaluated and potentially eliminated from the rating scale category structure. Critical judgments must be made in this process as the stepwise ordering between response categories must remain intact. For example, if an item demonstrates evidence of misfit in Category 2, it would warrant the consideration of collapsing (i.e., combining) into an adjacent category to maintain a stepwise ordering within the category structure.

There are multiple approaches to the collapsing of categories. The purpose of collapsing categories is to properly organize disordered thresholds. Linacre (2002) suggests that every response category should have at least 10 observations and that observations should be distributed somewhat evenly among the categories. Bond and Fox (2015) suggest only collapsing categories when it makes substantive sense. Items leaning in only one direction (Items 12, 13, 23) should also be carefully evaluated. Although the rating scale allows the investigator to determine the level of agreement or disagreement, it does not allow the dichotomous separation of ability. It is under the suggestions of Bond and Fox that categories were collapsed in this study, resulting in the revised rating scale (Figure S1 in the online version of the article).

Appendix D (in the online version of the article) shows a summary for the calibration of all raters. Rater severity ranged from 1.98 (Rater 13, most severe) to −1.78 (Rater 1, most lenient). Raters 3, 5, 7, 8, 9, 10, and 11 all have infit MSE values of less than 0.8, indicating muted response patterns. Rater 16, however, resulted in an infit MSE of 2.13, which indicates an irregular, or unexpected, response pattern.

Appendix E (in the online version of the article) displays a summary of the statistics for rater type. Administrators were placed on the logit scale at 0.41, while the music content experts were placed at −0.42. These measures indicate that overall, administrators demonstrate higher severity in scoring than music content experts. Infit and outfit MSE values fall within the required range (0.8–1.2), implying acceptability of fit in regards to the rater type.

Table 2 is a display of the differential rater functioning (DRF) statistics. DRF is exhibited when raters show systematic levels of severity or leniency among different subgroups (Engelhard, 2008). DRF is indicated by a *Z* score higher than 2.00 or below −2.00. This table shows specific items in which raters exhibited highly unexpected (overly lenient or severe) behavior. There were a total of 38 interaction terms, 5 of which are indicated by a Z score ±2.00. Specifically, Item 8 shows opposite behavior depending on the rater type. Administrators were far more lenient on Item 8, while music specialists were highly irregular in their ratings.

Overview of Results

The first research question investigated which items demonstrate acceptable model fit for the construct of lesson plan development in the setting of a secondary-level music performance classroom. Overall, the majority of items demonstrated good model fit (see Table 1 and Figure S1 in the online version of the article). However, a total of five items did not adequately fit the model. First, Item 8 read, “activities permit student choice.” Administrators and music specialists did not treat this item similarly, perhaps the reason why the item did adequately fit the model. Item 9 addressed the teacher’s

Table 2. Summary of Differential Rater Functioning Statistics (Rater Interactions) for Selected Raters exhibiting $|\mathbf{Z}| \geqslant 2.0$.

Item Number	Rater Type	Infit MSE	Outfit MSE	Total Observed	Total Expected	Standard Mean Residual (obs-exp)	Bias Logit	*SE*	*Z*
23	Music content specialist	0.90	1.10	100	95.15	0.15	1.11	0.45	2.47
19	Administrator	0.80	0.90	100	92.42	0.24	0.62	0.29	2.10
8	Music content specialist	1.60	1.70	72	63.89	0.25	0.58	0.26	2.24
8	Administrator	0.60	0.60	75	83.12	−0.25	−0.53	0.26	−2.06
23	Administrator	1.20	1.00	96	100.83	−0.15	−1.03	0.48	−2.12

Note. In the context of rater-mediated assessments, infit and outfit MSE statistics below 0.80 have been found to suggest "muted" ratings (i.e., possible dependencies), and values greater than 1.20 have been found to suggest "noisy" ratings (i.e., many unexpected observations) (Engelhard, 2013).

statement of connection to other disciplines. This item may not be explicitly stated within a lesson plan and may come more organically if a teacher were to be observed. Items 10 and 20 both addressed differentiation, which may be more difficult for pre-service teachers to explicitly state in a lesson plan. In addition, an administrator may overlook differentiation within a music classroom as a performance-based classroom looks different from a content-driven classroom. Item 18 addressed authentic learning through real-life examples. This item may not fit due to an unclear definition of authentic learning or a lack of transparency in providing students with a connection to the outside world. Assessment was addressed in Items 28 and 33. As pre-service teachers have reported feeling ill-prepared in this area, it can be concluded that clear plans for assessment are not seen within the lesson plan.

The second research question investigated how the rating scale changed structure as the raters showed inconsistent usage of particular categories. For the majority of the items in the revised rating scale (see Figure 1), only three out of four categories were used. In general, the categories most eliminated were the extremes of the categories, either *strongly disagree* or *strongly agree*. Only four cases existed where all four rating categories were used consistently (Items 11, 15, 19, and 26). In addition, four cases existed where two categories were eliminated (Items 12, 13, 23, and 34). Some items showed a general positive leaning for all raters (Items 12, 13, 23). The revised rating scale allowed for a more accurate evaluation of pre-service teachers' lesson plans as only applicable items and rating scale categories remained.

The final research question investigated the presence of differential rater severity between administrators and music specialists across items. This differential severity was present among three items only. First, Item 8 addressed student choice. Administrators were likely expecting a clear plan for students to make clear choices, while music specialists may not have expected student choice to be included in the

1. Develops plans that are clear.	Disagree	Agree		Strongly Agree
2. Develops plans that are logical.	Disagree	Agree		Strongly Agree
3. Develops plans that are sequential.	Disagree	Agree		Strongly Agree
4. Plans instruction effectively for pacing.	Disagree	Agree		Strongly Agree
5. Transitions are logical and sequential.	Disagree	Agree		Strongly Agree
6. Aligns and connects lesson objectives to standards.	Disagree	Agree		Strongly Agree
7. Develops appropriate daily plans.	Disagree	Agree		Strongly Agree
11. Organizes the lesson to progress toward a deep understanding of content (i.e. content mastery).	Strongly Disagree	Disagree	Agree	Strongly Agree
14. Reinforces learning goals throughout the lesson.	Disagree	Agree		Strongly Agree
15. Effectively uses appropriate instructional technology to enhance student learning.	Strongly Disagree	Disagree	Agree	Strongly Agree
16. Develops higher order thinking through questioning.	Strongly Disagree	Disagree		Agree
17. Encourages critical thinking through problem solving activities.	Strongly Disagree	Disagree		Agree
19. Teacher's plans reference curricular frameworks or blueprints to ensure accurate sequencing.	Strongly Disagree	Disagree	Agree	Strongly Agree
21. Provides remediation and enrichment to further student understanding of material.	Strongly Disagree	Disagree		Agree
24. Demonstrates high expectations for all students in content mastery.	Disagree	Agree		Strongly Agree
25. Plans follow-up activities designed to meet varied abilities of students.	Strongly Disagree	Disagree		Agree
26. Aligns student assessment with established objective.	Strongly Disagree	Disagree	Agree	Strongly Agree
27. Involves students in setting learning goals and monitoring their own progress.	Strongly Disagree	Disagree		Agree

(continued)

29. Uses formal assessments for diagnostic, formative, and summative purposes.	Strongly Disagree	Disagree	Agree
30. Uses informal assessments for diagnostic, formative, and summative purposes.	Disagree	Agree	Strongly Agree
31. Uses assessment techniques that are appropriate for the developmental level of students.	Disagree	Agree	Strongly Agree
32. Uses diagnostic assessment data to develop learning goals for students.	Strongly Disagree	Disagree	Agree
34. Uses diagnostic assessment data to document learning.	Disagree	Agree	

Figure 1. Revised 26-Item Lesson Plan Evaluation Rating Scale.

performance of music. Next, Item 19 addressed the reference to curricular frameworks and accurate sequencing. Administrators demonstrated severity on the rating of this item, likely because lesson plans did not explicitly state a sense of sequencing. Music specialists, who understood which skills are required to move from task to task, could see the sequencing without a clear statement from the pre-service teacher. Last, Item 23 referred to planning at an appropriate content level.

Discussion

Pre-service music educators have reported feeling inadequately prepared for teaching in the performing arts classroom and especially have a perceived lack of understanding of teacher evaluation (Duncan, 2011). In addition, pre-service teachers have reported the need for greater attention during preparation programs in the areas of music curriculum, lesson planning, and student assessment (Berg & Miksza, 2010; Conway, 2002b; Snyder, 1998). A multistep process was employed to develop an instrument that could evaluate students' lesson plans. After creating the observational design, a Likert-type scale was used for raters to evaluate each item. The judging plan was formed and rater data collected. When analyzing data, misfit is extremely sensitive and must be handled with careful scrutiny. Once the misfit was managed, the original measure had to be refined (Figure 1). Finally, the rating scale structure was evaluated and optimized.

As discussed earlier, good model-data fit and high reliability of separation indicate a strong argument for construct validity. Therefore, any alterations to the rating scale itself include the elimination of misfit items and changes to the rating scale category structure. We acknowledge the divergence of response based on the type of rater (e.g., academic administrators vs. music education content specialists). Ideally, both administrators and music specialists would undergo some sort of rater training protocol to

align their ratings to fair and equitable rating practices. Music specialists would likely have a more accurate understanding of what type of planning is most appropriate for each level of teaching. Particular musical training for administrators would likely aid in this assessment. However, these types of training most likely are not feasibly due to challenges related to time, money, standards, and so on. Although it is not ideal, we can control for rater differences in the measurement model itself to the best of our ability.

Face validity is a qualitative, contextual way to approach the validity of a rubric. Empirical data, however, are the best way to examine construct validity. With face validity, items cannot be added or dropped based on perceptions. To substantiate such processes, the rubric must be revalidated with empirical evidence of how well the overall construct functions. The revised instrument in this study does not maintain consistent categories among items. The decisions leading to the inclusion of specific categories should be empirically based. The instrument gives us the strongest interpretation of the function of items and rating scale categories that ultimately defines the construct. Any changes made to the instrument due to face validity would be speculation at best and were not considered as part of this study. A future revalidation study to include considerations of the perceptions and use of the final instrument resulting in this study is therefore suggested.

As pre-service teachers transition from the college setting to the classroom setting, a significant shift occurs as evaluation moves from the hands of music education specialists to school administrators. These administrators often have teaching backgrounds in non–performing related subjects, such as language arts or social studies. More specifically, administrators in this study came from backgrounds in career and technical education, counseling, language arts, mathematics, foreign language, science, and social studies. This study suggests that academic administrators are, overall, more severe evaluators than music education content specialists. This gap in severity could stem from the lack of content-specific knowledge by the administrator. The gap in music teaching expectations may also stem from the location of music teacher preparation programs. In some colleges and universities, preparation for music teachers is housed in a department or school of music. Preparation for teachers of other subjects, such as science, mathematics, and social studies, is housed in colleges or departments of education. These pre-service teachers are prepared under a common set of standards and expectations that pre-service music teachers may not encounter. However, these are speculative considerations and warrant further phenomenographic investigations.

The discrepancy in the expectations of administrators and music education professors can challenge young teachers to first be more explicit in their lesson plans. Much of the jargon used in music teaching is foreign to non–music educators. In non-arts disciplines, more familiar techniques of differentiation, remediation, and enrichment can all be observed as an administrator moves through the room. In a music classroom, these evaluation components are frequently used without the direct indication in a lesson plan. Administrators could be better trained on how to look for these components within different types of classrooms.

In-service teacher evaluation procedures focus on the improvement of teaching behaviors and overall student learning. However, these procedures may have other

consequences. Nelson (2012) discusses teachers' concerns about the additional time and work needed to prepare for the evaluation process. Stresses related to evaluation may impact teacher retention. In addition, some states use teacher evaluation as a means to determine teacher salary. Because of the wide variety of expectations for teacher effectiveness, teacher evaluation might look different between counties and states (Nelson, 2012). There is also a need for evaluation systems to be constantly revalidated to ensure accurate outcomes.

The adoption of the Common Core curriculum has directly impacted teacher accountability. Classroom teachers are charged with increased responsibility to document student learning specifically in the form of individual growth. The need to document student growth suggests that pre-service preparation should include more in-depth assessment strategies. Potentially, administrators may view this integration of Common Core into the arts curriculum as a need for project-based learning (Taylor, 2014). Teachers are being asked to go beyond the daily lesson and rehearsal. Pre-service teachers should approach this not as a complication but as a way to provide engaging activities for every type of learner in the classroom.

The future of teacher preparation should be an integration of expectations from both administrators and music education specialists. Music education professors are ensuring that pre-service teachers understand content-specific skills and can effectively impart knowledge to future music students. Administrators, on the other hand, are more concerned with overall student learning and growth. Their jobs rely on teacher effectiveness through successful teacher and student evaluation.

Teacher effectiveness directly impacts student success. Many teacher preparation programs focus dually on content-specific material as well as teaching strategies. Teaching strategies refers to the variety of instruction given in the classroom and should be manipulated based on how students learn. Content-specific material refers to the presentation and understanding of music-related content. Although a pre-service teacher may have mastered the music content, he or she may not be able to present that material in a way that meets the needs of the students. The gap between content-specific material and teaching strategies occurs when pre-service teachers are not given ample time in the public school before teaching. Pre-service teachers also need more time to understand the requirements of public school administrators. In addition, administrators need to understand the inner workings of the music classroom. More training should be provided to administrators to recognize differences in instruction while moving from an academic classroom to a performance-based classroom. If more consistency can be provided as pre-service teachers transition from college to the public school classroom, teachers will be set up for more success.

The revised Lesson Plan Rating Scale provides an opportunity to help pre-service teachers become more familiar with classroom expectations. By incorporating this Rating Scale into the curriculum, pre-service teachers will have the potential to provide more comprehensive lesson plans to evaluate and promote student learning within the music classroom. This Rating Scale should go through the revalidation process to ensure accurate outcomes and then should be transferred to a rubric format for implementation in the classroom. Therefore, constant use and monitoring will only aid in its

ability to provide accurate feedback to the user. This Lesson Plan Rating Scale serves as a way to communicate between a teacher's expectations and a student's performance and should be used to further the discussion of quality teaching.

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.

Supplemental Material

Appendices A through D and Figures S1 through S4 are available in the online version of the article at https://doi.org/10.1177/0022429418793645

References

Akyuz, D., Dixon, J. K., & Stephan, M. (2013). Improving the quality of mathematics teaching with effective planning practices. *Teacher Development*, *17*(1), 92–106. doi:10.1080/13664530.2012.753939

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145–1146.

Berg, M. H., & Miksza, P. (2010). An investigation of preservice music teacher development and concerns. *Journal of Music Teacher Education*, *20*(1), 39–55. doi:10.1177/1057083710363237.

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

Brittin, R. V. (2005). Preservice and experienced teachers' lesson plans for beginning instrumentalists. *Journal of Research in Music Education*, *53*, 26–39. doi:10.1177/002242940505300103

Butler, A. (2001). Preservice music teachers' conceptions of teaching effectiveness, microteaching experiences, and teaching performance. *Journal of Research in Music Education*, *49*, 258–272. doi:10.2307/3345711

Butt, G. (2006). *Lesson plan* (2nd ed.). London: Continuum International Publishing Group.

The Center for Educator Effectiveness, McREL International. (2013). *McREL's research-based teacher evaluation system: The CUES framework*. Denver, CO: McREL International.

Chaffin, C. R. (2009). Perceptions of instrumental music teachers regarding the development of effective rehearsals. *Bulletin of the Council for Research in Music Education*, *181*, 21–36.

Conway, C. (2002a). Curriculum writing in music. *Music Educators Journal*, *88*(6), 54–59. doi:10.2307/3399806

Conway, C. (2002b). Perceptions of beginning teachers, their mentors, and administrators regarding preservice music teacher preparation. *Journal of Research in Music Education*, *50*, 20–36. doi:10.2307/3345690

Coppola, A. J., Scricca, D. B., & Connors, G. E. (2004). *Supportive supervision: Becoming a teacher of teachers*. Thousand Oaks, CA: Corwin Press.

Council for the Accreditation of Educator Preparation. (2016). *2013 CAEP standards*. Retrieved from http://caepnet.org/standards/introduction

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision & Curriculum Development.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: The Danielson Group.

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

Duncan, A. (2011). *Our future, our teachers: The Obama administration's plan for teacher education reform and improvement*. Retrieved from https://www.ed.gov/sites/default/files/our-future-our-teachers.pdf

edTPA, Annual Administrative Report. (2015). *Educative assessment and meaningful support*. Stanford, CA: Stanford Center for Assessment, Learning, and Equity.

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

Engelhard, G. (2008). Differential rater functioning. *Rasch Measurement Transactions*, *21*(3), 281–385.

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

Engelhard, G., Jr., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research & Perspective*, *9*, 40–45. doi:10.1080/15366367.2011.558787

Frey, N., Fisher, D., & Moore, K. (2005). *Designing responsive curriculum: Planning lessons that work*. Lanham, MD: Rowman & Littlefield Education.

Furr, R. M., & Bacharach, V. R. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: SAGE Publications.

Houston, D., & Beech, M. (2002). *Designing lessons for the diverse classroom: A handbook for teachers*. Retrieved from http://www.fldoe.org/core/fileparse.php/7690/urlt/0070084-4dclessn.pdf

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lane, J. S. (2006). Undergraduate instrumental music education majors' approaches to score study in varying musical contexts. *Journal of Research in Music Education*, *54*, 215–230. doi:10.1177/002242940605400305

Lane, J. S., & Talbert, M. D. (2015). Examining lesson plan use among instrumental music education majors during practice teaching. *Journal of Music Teacher Education*, *24*(3), 83–96. doi:10.1177/1057083713514979

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. (1989). *Many-facet Rasch measurement*. Chicago, IL: Mesa Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

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

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

Linacre, J. M. (2017). *Reliability and separation of measures*. Retrieved from http://www.winsteps.com/winman/reliability.htm

Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many-facet data. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theories, models, and applications* (pp. 296–321). Maple Grove, MN: JAM Press.

"Linking terminology: Raw score and Rasch." (2007). *Rasch Measurement Transactions*, *20*(4), 1076.

Marzano Research Laboratory. (2013). *The Marzano teacher evaluation model*. Englewood, CO: Marzano Research Laboratory. Retrieved from http://tpep-wa.org/wp-content/uploads/Marzano_Teacher_Evaluation_Model.pdf

Marzano, R. J. (2003a). *Classroom management that works*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2003b). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2006). *Classroom assessment and grading that work*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works*. Alexandria, VA: Association for Supervision and Curriculum Development.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Miksza, P., & Berg, M. H. (2013). Transition from student to teacher: Frameworks for understanding pre-service music teacher development. *Journal of Music Teacher Education*, *23*, 10–26. doi:10.1177/1057083713480888

National Association of Schools of Music. (2016). *National Association of Schools of Music handbook*. Retrieved from https://nasm.arts-accredit.org/wp-content/uploads/sites/2/2015/11/NASM_HANDBOOK_2016-17.pdf

National Education Association. (2011). *New policy statement on teacher evaluation and accountability*. Retrieved from http://www.nea.org/grants/46326.htm

Nelson, J. A. (2012). *Effects of teacher evaluations on teacher effectiveness and student achievement* (Unpublished master's thesis). Northern Michigan University, Marquette, Michigan.

Paul, S. J. (1998). The effects of peer teaching experiences on the professional teacher role development of undergraduate instrumental music education majors. *Bulletin of the Council for Research in Music Education*, *137*, 73–92.

Schleuter, L. (1991). Student teachers' preactive and postactive curricular thinking. *Journal of Research in Music Education*, *39*, 48–65. doi:10.2307/3344608

Schmidt, M. (2005). Preservice string teachers' lesson-planning processes: An exploratory study. *Journal of Research in Music Education*, *53*, 6–25. doi:10.1177/002242940505300102

Shuler, S. C., Brophy, T. S., Sabol, F. R., McGreevy-Nichols, S., & Schuttler, M. J. (2015). Arts assessment in an age of accountability: Challenges and opportunities in implementation, design, and measurement. In H. I. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 183–216). New York, NY: Routledge, Taylor & Francis Group.

Snyder, D. W. (1998). Classroom management for student teachers. *Music Educators Journal*, *84*(4), 37–40. doi:10.2307/3399115

Stronge & Associates. (2013). *Stronge teacher evaluation system: A validation report*. Retrieved from http://www.cesa6.org/effectiveness_project/Validation-Report-of-Stronge-Evaluation-System.pdf

Stronge, J. H., Ward, T., & Xu, X. (2013). *Virginia teacher evaluation and Virginia Performance-Pay Incentives (VPPI) pilot: An evaluation report*. Richmond, VA: Virginia Department of Education.

Taylor, P. (2014). *Integrating arts learning with the common core state standards*. Retrieved from http://ccsesa.org/wp-content/uploads/2014/12/FINAL-Common-Core-Publication.compressed.pdf

Teachout, D. J. (1997). Preservice and experienced teachers' opinions of skills and behaviors important to successful music teaching. *Journal of Research in Music Education*, *45*, 41–50. doi:10.2307/3345464

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

Virginia Department of Education. (2012). *Guidelines for uniform performance standards and evaluation criteria for teacher*. Retrieved from http://www.doe.virginia.gov/teaching/performance_evaluation/teacher/index.shtml

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*(1), 77–85.

Wesolowski, B. C. (2015). Tracking student achievement in music performance: Developing student learning objectives for growth model assessments. *Music Educators Journal*, *102*(1), 39–47.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception*, *33*, 662–678.

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

Woods, R. (2015). *Teacher keys effectiveness system: Implementation handbook*. Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TKES&20Handbook%20-713.pdf.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Author Biographies

Dorothy J. Musselwhite received her PhD from the University of Georgia. Her research interests include assessment and pre-service teachers.

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

Submitted February 2, 2017; accepted November 21, 2017.