THE PSYCHOMETRIC EXAMINATION OF PRE-SERVICE MUSIC EDUCATORS' QUALITY OF VERBAL FEEDBACK IN THE SECONDARY-LEVEL INSTRUMENTAL MUSIC CLASSROOM

by

MYRIAM I. ATHANAS

(Under the Direction of Brian C. Wesolowski)

ABSTRACT

The quality of verbal feedback teachers provide to students in the instrumental music classroom is directly related to students' real-time performance and execution of rehearsed music. The purpose of the first study (see Chapter 2) was to examine the quality of pre-service music educators' verbal feedback in the context of secondary-level instrumental ensemble rehearsals through the development and validation of a verbal feedback evaluation scale. The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* is based on a four-point Likert-type scale including 35 criteria embedded within five instructional domains. Implications for quality teaching and rehearsals, as well as pre-service music education training in the secondary-level instrumental classroom are discussed.

The purpose of the second study (see Chapter 3) was to examine the self-assessment accuracy of pre-service music educators' quality of verbal feedback in the context of secondary-level instrumental ensemble instruction. The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* was used to examine the accuracy between content experts' evaluation of

students and students' self-evaluation for the same teaching episode at both item and domain levels.

The purpose of the third study (see chapter 4) was to determine if pre-service music educators could be grouped by common patterns into distinct typologies based on the quality of their verbal feedback in the music ensemble rehearsal. Considerations for the inclusion of self-assessment accuracy measures in teacher preparation curricula and its role in improving student-teacher communication, instructional quality, differentiated instruction, and reflective practice are discussed.

INDEX WORDS: Accuracy, Assessment, Differentiation, Instructional measurement, Pre-service, Rasch, Rating scale, Self-assessment, Verbal feedback

THE PSYCHOMETRIC EXAMINATION OF PRE-SERVICE MUSIC EDUCATORS' QUALITY OF VERBAL FEEDBACK IN THE SECONDARY-LEVEL INSTRUMENTAL MUSIC CLASSROOM

by

MYRIAM I. ATHANAS

B.Mus., Kennesaw State University, 2013

M.M.E., The University of Georgia, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

© 2021

Myriam I. Athanas

All Rights Reserved

THE PSYCHOMETRIC EXAMINATION OF PRE-SERVICE MUSIC EDUCATORS' QUALITY OF VERBAL FEEDBACK IN THE SECONDARY-LEVEL INSTRUMENTAL MUSIC CLASSROOM

by

MYRIAM I. ATHANAS

Major Professor:	Brian Wesolowski
Committee:	Alison Farley
	George Engelhard
	Peter Jutras

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

2 THE DEVELOPMENT OF A SCALE TO MEASURE THE QUALITY OF PRE-SERVICE TEACHERS' VERBAL FEEDBACK IN THE CONTEXT OF SECONDARY-LEVEL INSTRUMENTAL MUSIC EDUCATION 5

Abstract 6

Introduction and Literature Review 7

Method 16

Results 21

Conclusion 25

Discussion 26

3 EXAMINING SELF-ASSESSMENT ACCURACY OF PRE-SERVICE MUSIC EDUCATORS' QUALITY OF VERBAL FEEDBACK 42

Abstract 43

Introduction and Literature Review 44

Method 51

Results 56

Conclusion 60

Discussion 61

4 ATTRIBUTES OF PRE-SERVICE MUSIC EDUCATORS' VERBAL FEEDBACK IN THE SECONDARY-LEVEL INSTRUMENTAL MUSIC CLASSROOM 72

Abstract 73

Introduction and Literature Review 75

Method 80

Results 83

Conclusion and Discussion 88

5 CONCLUSIONS 97

REFERENCES 102

LIST OF TABLES

Page

Table 1.1: Summary Statistics from the Many Facet Rasch Partial Credit Model 29

Table 1.2: Calibration of Domain and Criteria Facets 30

Table 1.3: Category Diagnostics: Category Usage, Average Observed/Expected Measures, and Outfit MSE 32

Table 1.4: Calibration of Student Facet 34

Table 1.5: Calibration of Rater Facet 36

Table 2.1: Summary Statistics from the Multifaceted Rasch Rater Accuracy Model 63

Table 2.2: Calibration of Student Facet 64

Table 2.3: Calibration of Criteria 66

Table 2.4: Calibration of Domains 67

Table 3.1: Finalized Cluster Centroids by Scale Criteria 90

LIST OF FIGURES

Page

Figure 1.1: Pre-service Music Teacher Verbal Feedback Evaluation Scale 37

Figure 1.2: Scale Items to be Removed or Reexamined 40

Figure 1.3: Wright Map for the Many Facet Rasch Partial Credit Model 41

Figure 2.1: Pre-Service Music Teacher Verbal Feedback Evaluation Scale (Revalidated) 68

Figure 2.2: Variable Map for the Many Facet Rasch Rater Accuracy Model 71

Figure 3.1: Pre-service Music Teacher Verbal Feedback Evaluation Scale 91

Figure 3.2: Contribution of Scale Criteria to Dimension 1 and Dimension 2 from the Principal Components Analysis 94

Figure 3.3: Contribution of all Scale Criteria in the Two-Dimensional Space 95

Figure 3.4: Three-cluster Solution for Possible Quality Verbal Feedback Typologies 96

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Historically, music teacher educators have struggled to decide whether the standard 4-year undergraduate program, which allows students to become certified to teach a broad range of specialties upon completion of their degree programs, is sufficient preparation for pre-service teachers to enter the classroom and be successful music educators (Hash, 2020). Undergraduate music education majors are asked to learn a tremendous amount of information in a limited amount of time; however, a conflict may lie in *what* and *how* information is presented to students, not solely on the amount of *time* pre-service teachers have to master the curricular content (Yarbrough et al., 1979). A sizeable portion of the curriculum in the undergraduate music education program of study, although an important component, focuses on music content (e.g., music theory, aural skills, and music history) and developing satisfactory performance skills. Music education majors' curricular studies are comprised of nearly 30% more credited course hours (i.e., courses taken for greater than zero credit hours) labeled as "content knowledge" than courses focused on developing and refining the pedagogical and praxis skills required for students to become exceptional educators (Haning, 2021). Learning subject matter content and learning to become an "accomplished teacher" are disparate, and learning to teach is a long-lasting and intentional developmental process involving a variety of experiences (e.g., reflective practices to inform one's teaching, adaptive abilities, assessment, etc.) that foster pre-service teacher maturation (Feiman-Nemser, 2008). Therefore, a proportionate amount of the music curriculum should be spent on teaching-specific content as on general music content.

Pre-service as well as experienced in-service music educators have indicated that they valued their fieldwork, student teaching, and sometimes their instrumental methods courses, but believed that the lack of context and cohesiveness in which information was presented in their university courses caused them to perceive their experiences as less constructive, practical, and lacking in relevance (Conway, 2002, 2012). Traditional teacher preparation programs are often criticized for the apparent disconnect between theory and practice, as learning does not happen in the university classroom *or* the school classroom alone, and therefore involves a combination of both settings in order to provide a variety of experiences to study pedagogy, research, practice, and self-assessment (Gonzo & Forsythe, 1976; Darling-Hammond, 1999). Choy et al. (2014) examined the perceptions pre-service teachers had on the relevance of the courses in their major programs and how their programs prepared them for practicum experiences. Researchers found that the student teachers felt the first two years of their major coursework was solely focused on content and had little relevance to teaching practices, foundations of education, nor helped prepared them to face the challenges of real-world classroom teaching. When pre-service teachers enter the classroom, they realize that they cannot apply what they have learned in their methods courses into an authentic teaching context and that is an inherent problem with how they are being taught in their teacher training programs (Bannister & Linder, 2015). Simple exposure to the theoretical knowledge of teaching practices is not enough and students need to put their theory into action in authentic contexts (Sadler, 2010).

Asmus (2000) suggested that, "Music teacher education has never before needed a base of substantive information about how best to prepare music teachers as it does now…[because] simply put, the days of the general music, band, orchestra, and chorus foci are over" (p. 5). This statement still rings true today. The lack of a comprehensive assessment curriculum may come

from the absence of a knowledge base surrounding what quality assessment looks like in pre-service teacher education (Deluca & Klingerb, 2010). Education measurement theory has made huge strides over the last century (Hattie, Jaeger, & Bond, 1999). Modern measurement theories founded in educational measurement, such as Item Response Theory (Wesolowski, 2019), are consistently being utilized in the field of music education to ensure that performance assessments are valid, reliable, and fair (Edwards, Edwards, & Wesolowski, 2018; Edwards, Edwards, & Wesolowski, in press; Musselwhite & Wesolowski, 2018; Musselwhite & Wesolowski, 2021; Wind & Wesolowski, 2018; Wesolowski, 2016; Wesolowski et al., 2016; Wesolowski et al., 2017; Wesolowski et al., 2017; Wesolowski et al., 2018). We must prepare our pre-service teachers for an empirical data-driven society by cultivating a fundamental understanding of what high-quality (i.e., valid and reliable) formative and summative assessment methods look like and how to utilize them to inform teaching and help students achieve their specific learning goals (Kaschub & Smith, 2014).

Formative assessment through verbal feedback in the performance-based classroom is a cyclical process whereby information is exchanged multiple times between the teacher and the student in order to create a student-centered culture of learning and to cultivate problem solving practices students can continue to use once they leave a teacher's classroom (Mulliner & Tucker, 2017). Giving immediate verbal feedback in a rehearsal setting is a stressful process in which music educators are required to make qualitative human judgements in order to guide students through a meaningful and methodical learning process (Sadler, 1998). Teachers cannot expect to have a skill they have not practiced in authentic contexts and cannot be expected to administer appropriate verbal feedback when they have not been repeatedly exposed to authentic teaching situations where they are required to judge the quality of a student's work and give them

feedback in an effective and efficient manner (Sadler, 2010). This dissertation consists of three manuscripts examining the psychological construct of the quality of verbal feedback in the secondary instrumental ensemble classroom. The intention of these articles is: (a) to provide an empirical formative assessment tool which can evaluate teachers' verbal feedback in a valid and reliable way and is easily accessible to pre-service and in-service music educators, (b) to assess whether pre-service music educators have the ability to accurately assess their own teaching when their self-assessment ratings are compared with the ratings of a teacher education content expert, and (c) to use the empirical data gathered from this formative assessment of teachers' verbal feedback to enact change within teacher training programs and courses. As will be thoroughly discussed and emphasized throughout this document, information gathered from formative assessments should be used to inform teaching practices as well as lesson design and implementation for a diverse array of learners in secondary music classrooms (Burrack and Parkes, 2019). Pre-service teacher education carries serious implications which support the need for rigorous teacher preparation programs that are not solely focused on the knowledge of subject matter, but also a deep understanding of how to teach that information in a pedagogically sound and relevant way to their future students (Howard & Aleman, 2008).

CHAPTER 2

THE DEVELOPMENT OF A SCALE TO MEASURE THE QUALITY OF PRE-SERVICE TEACHERS' VERBAL FEEDBACK IN THE CONTEXT OF SECONDARY-LEVEL INSTRUMENTAL MUSIC EDUCATION[1]

[1] Athanas, M.I. and B.C. Wesolowski. To be submitted to *Bulletin of the Council of Research in Music Education.*

Abstract

The purpose of this study was to examine the quality of pre-service music educators' verbal feedback in the context of secondary-level instrumental ensemble rehearsals through the development and validation of a pre-service music teacher verbal feedback evaluation scale. The questions that guided this study include: (a) What are the psychometric qualities (e.g., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale? (b) How do the verbal feedback criteria vary in difficulty in relation to how the students vary in achievement? and (c) How does the rating scale category structure vary across each individual criterion? A sample of pre-service music educators' teaching episodes ($N = 55$) was video recorded in 6–10-minute segments. Music content experts ($N = 15$) evaluated the teaching segments using the Pre-Service Music Teacher Verbal Feedback Evaluation Scale consisting of 39 criteria embedded within five domains. Data were analyzed using the Many Facet Rasch Partial Credit (MFR-PC) model. Results indicated a high reliability of separation and a good data-to-model fit for the MFR-PC. Implications for teaching and rehearsal effectiveness as well as pre-service music education training in the secondary-level instrumental classroom are discussed.

Keywords: assessment, instructional feedback, pre-service, rating scale, Rasch

Introduction

In an instrumental music ensemble rehearsal, music teachers' formative, verbal feedback is an important mechanism for providing real-time performance assessments of students. Formative feedback is defined as "information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning" (Shute, 2008, p.154). Music teachers provide formative feedback frequently in a music rehearsal and according to Duke (2012), "expert teachers, throughout a learning sequence, make many evaluative statements concerning the quality of students' performances moment to moment" (p. 132). Formative feedback should be descriptive, positive, and constructive as it is an important method in performance-based learning to correct student errors as they happen in real-time. This type of feedback, particularly in the context of a music ensemble rehearsal, is the most frequent impetus toward improving the performance achievement of individual students, small groups of students, or the ensemble as a whole (Hale & Green, 2009; Brookhart, 2017). In music ensemble rehearsals, students inevitably make performance errors when learning new skills. However, students can efficiently improve when provided adequate and sufficient formative feedback from the teacher. For any music teacher, an instructional challenge is to fluently and efficiently deliver adequate and immediate formative feedback to students throughout the music rehearsal in real-time.

Music instruction is complex in nature, particularly due to the challenge of making continuous and immediate aural evaluations of student performances in real-time (Brand & Burnsed, 1981; DeCarbo, 1984). Nevertheless, music educators are expected to aurally evaluate performance errors and administer feedback on student performances, consider pedagogical strategies for their improvement, and communicate remedial instructions all while concurrently

engaging with students (Saunders & Holahan, 1997). In the field of music education, multiple measures are available to deliver quality, written summative feedback to both individual students (Abeles, 1973; Bergee, 1988; Ciorba & Smith, 2009; Jones, 1986; Saunders & Holahan, 1997, for example) and groups of students (Cooksey, 1977; Russell, 2010; Smith & Barnes, 2007; Smith, 2009; Zdzinski & Barnes, 2002, for example). Although these tools allow us to administer quality summative feedback to students, the field is limited in the ability to systematically evaluate the quality of music teachers' formative verbal feedback. The development and evaluation of a measure to assess the quality of music teachers' formative verbal feedback (i.e., real-time, moment to moment, formative assessments) has not been systematically examined. Orrell (2006) suggests that, "providing students with focused, comprehensive feedback on their learning product is a significant aspect of teaching and assessing" (p. 443). Furthermore, as Colwell (1999) suggests, "I see students who receive no immediate constructive feedback failing. (*Constructive* is an important word here; music students do receive gratuitous, unearned group praise)" (p. 33). For pre-service and less experienced music teachers, specifically, administering high-quality, effective, and meaningful feedback is a skill they often struggle to develop (Duke, 2012). Therefore, we decided to use pre-service music teachers as the subjects of this study due to the importance of developing the quality of their formative feedback.

Verbal Feedback and Formative Performance Assessment

The quality of verbal feedback teachers provide to students in the instrumental music classroom is directly related to students' real-time performance and execution of the rehearsed music (Duke & Madsen, 1991). Additionally, the verbal feedback students receive through quality teacher interventions during the learning process can impact both the way students view their ability to perform the given musical task and their ability to set and modify realistic musical

goals (Senko & Harackiewicz, 2005). The ability of a student to modify learning goals after receiving teacher feedback is referred to as "goal switching." According to Senko and Harackiewicz:

> … performance feedback may change a student's perceived competence and, consequently, the student's further pursuit of an achievement goal as well…. we propose that students might regulate their achievement goal pursuit after receiving early positive or negative competence feedback that allows them to evaluate their progress toward the goal (p. 321).

Certain characteristics of quality teacher feedback, including the effects of feedback by modeling (Rutkowski & Miller, 2003), accuracy of positively and negatively perceived feedback (MacLeod & Nápoles, 2013), and types of applied feedback (Cranmore & Wilhelm, 2017) were systematically examined in music education research. However, these considerations only cover a small portion of criteria that can be used to define the overall quality of verbal feedback teachers deliver to students. Furthermore, detailed criteria for what constitutes quality, formative verbal feedback in a music rehearsal is limited in the research literature. Therefore, we examined both music and other performance-based academic fields to identify criteria that may constitute high quality, formative verbal feedback in the context of a secondary-level music rehearsal.

Defining Quality in Formative Feedback: Domain Considerations

An extensive review of the educational research literature in music education and other performance-based academic fields (e.g., English and Language Learning) were examined in order to identify key elements of high-quality verbal feedback. Five common themes emerged relevant toward secondary-level music ensemble instruction: (a) context/audience of feedback, (b) learning objective/focus of feedback, (c) timing of feedback, (d) type of feedback, and (e)

tone of feedback. The following provides a brief overview of each theme and considerations for their importance toward improving the quality of music teachers' formative, verbal feedback.

Context/Audience of Feedback

In order for verbal feedback to be clear to students, the information should be appropriate for the teaching and learning contexts. The context of teacher feedback is defined by the classroom experience at hand (Brookhart, 2017). Examples of context include attributes related to (a) student characteristics; (b) the performance setting; and (c) the atmosphere of the music classroom.

Student characteristics in the music classroom are largely related to the audience (e.g., individual student or group) for which the feedback is being administered. In a music performance classroom (i.e., concert band, orchestra, or choral rehearsal settings), music educators provide individual-based feedback (i.e., feedback targeted at an individual student), small-group feedback (i.e., feedback targeted at smaller subsets of students), and large-group feedback (i.e., feedback targeted at the ensemble as a whole). Tindale et al. (1991) investigated the effects of individual verbal feedback versus group verbal feedback on individual students' achievement of music performance tasks in both positive (i.e., a successful performance where positive feedback is given) and negative (i.e., an unsuccessful performance where negative/constructive feedback is given) contexts. Results suggested a significant improvement in the performance of the individual student after receiving verbal feedback in the context of both group and individual settings when appropriate feedback is provided. This suggests it is important for the teacher to determine who the appropriate audience is (i.e., individual or group) when delivering verbal feedback. If feedback is perceived as irrelevant to the learning target, it is often dismissed as unimportant by the learner (Hattie & Timperley, 2007).

The *performance setting* refers to the particular musical ensemble that the teacher happens to be rehearsing (i.e., full ensemble, homogeneous/heterogeneous ensembles, small ensembles). Not all administrations of verbal feedback are appropriate for every performance setting. For example, Duke and Byo (2011) argue that group feedback is often irrelevant for at least one member of the ensemble. This argument emphasizes the importance of the setting in which verbal feedback is being delivered, so the teacher is able to effectively foster student improvement.

The *atmosphere* of the classroom refers to, "… a composite of variables working together to promote learning in a comfortable environment in a classroom" (Falsario et al., 2014). Variables in the instrumental performance setting may include the teacher/student relationship during rehearsal, the teacher's classroom expectations, behavior management, pacing, or student engagement throughout a rehearsal, for example. A positive classroom atmosphere has a welcoming climate that engages students in learning and helps to foster high values of self-efficacy (Hattie & Timperley, 2007). The use of positive feedback and praise (e.g., praising the student for success and giving positive feedback on why they were successful) is a tool that may help the teacher to create a positive classroom environment (Conroy et al., 2009). Duke (2012) suggests that it is the responsibility of the teacher to find a balance between positive and negative feedback. He emphasizes the importance for students to be both rewarded by being given a task they can be successful on (achievement at an independent-working level) and challenged by tasks that need more work (constructive feedback at a student's frustration-working level; Gickling & Thompson, 1985). He writes, "The teaching of experts is characterized by high rates of both positive and negative feedback," and that expert teachers, "control the rates of positive and negative feedback…by directing the tasks students perform so that the quality of

performance is predictable" (Duke, 2012, p. 133). A balance of positive and negative feedback can help to create a positive classroom atmosphere. It is important for students to feel comfortable receiving feedback about their mistakes within the classroom environment; therefore, a positive classroom atmosphere is an important factor in successful student learning (Conroy et al., 2009).

Focus of Feedback/Learning Objectives

The focus of verbal feedback refers to how teachers relate the quality of student work to the learning objectives underscoring the instructional episode (Perpignan, 2003). For feedback to be effective, it must be paired with an intended learning objective (Hattie & Timperley, 2007). Teachers may use verbal feedback as a formative assessment method to indicate to students to what degree they are meeting either the intended educational or instructional objectives. In order to facilitate students in meeting the intended objectives, the teacher can provide information specific to the learning process itself (e.g., the steps taken to achieve a specified goal) or about the specific student error committed. Duke and Madsen (1991) encourage teachers to be "proactive" in their feedback administration rather than "reactive" in order to provide opportunities that give meaningful and purposeful feedback to their students about the planned learning objective.

Timing of Feedback

The timing of verbal feedback refers to the teacher's choice to administer either immediate or delayed information to an individual student or group of students in order to help correct students' errors in real-time (Shute, 2008). Language Learning is an important performance-based academic area where the field of music can draw formative feedback criteria (Jordan-DeCarbo, 1986). Ellis (2009) synthesized a large volume of Language Learning

research, suggesting that immediate feedback is more effective than delayed feedback. Teachers' use of immediate feedback has shown to improve student learning and performance (Shute, 2008). Brookhart (2017) highlights some important points about the appropriate timing of verbal feedback:

> Feedback needs to come while students are still mindful of the topic, assignment, or performance question. It needs to come while they still think of the learning goal as a learning goal- that is, something they are still striving for, not something they already did. It especially needs to come while they still have some reason to work on the learning target. (p. 10-11)

The timing of verbal feedback is critical to the quality of the learning experience (Brookhart, 2017). Teachers' verbal feedback follows in the window of time immediately after student error, while the student is still attentive and mindful of the mistake to be corrected (Doughty, 2001). During this window of time, the teacher can provide valuable information about aspects of a student's classroom performance that they have observed (Hattie & Timperley, 2007). Another important criterion considered when constructing the scale was that students are offered time to apply teacher feedback (Sadler, 1983). Students require opportunities to evaluate, clarify, and apply feedback to their own learning processes.

Type of Feedback

Type of feedback refers to the category of the verbal feedback a teacher provides their students. Brunning et al. (2011) proposes two categories of verbal feedback: (1) information-oriented feedback (e.g., teacher provides students with feedback based upon a set of criteria to help them improve their performance), and (2) performance-oriented feedback (e.g., teacher compares one student's performance to another's). More explicitly, Brookhart (2017) proposes

three types of verbal feedback: (1) criterion-referenced feedback (e.g., comparing student performance to a set of criteria), (2) self-referenced feedback (e.g., comparing a student's performance to their own past performances), and (3) norm-referenced feedback (e.g., comparing student performances to each other). Her research suggests that criterion-referenced feedback and self-referenced feedback are more effective than norm-referenced feedback. Therefore, the focus of verbal feedback should be constructive and should highlight the strengths and weaknesses of a given performance based upon a pre-established standard (Brookhart, 2017). Ineffective verbal feedback (i.e., feedback that is not constructive, criterion-referenced or self-referenced in nature) may cause students to digress in their capabilities and become a deterrent to the learning process (Brookhart, 2008). Criterion-referenced and self-referenced are the two types of high-quality verbal feedback outlined in this study as they may encourage more meaningful and constructive feedback (Sadler, 1983).

Tone of Feedback

Tone is defined as the expressive quality of the feedback message (Brookhart, 2017). The tone of teacher verbal feedback can affect the way a student hears, receives, internalizes, and interprets the verbal feedback (Russell, 2009). The tone of a message is conveyed by a teacher's word choice (i.e., specific word uses which can elicit either positive or negative emotions from the learner) and style (i.e., the teacher's choice of using a question or statement to evoke a variety of student response types).

The directionality/connotation in which the feedback is delivered (i.e., positive or negative) is equally important to the quality of the verbal feedback. Chen, et al. (2011) suggest that the manner in which students perceive the tone or connotation of verbal feedback is a significant factor in predicting student learning achievement. Specifically, the perceptions of the

students' learning environment, as they relate to the teacher's use of positive tone quality when administering both positive and negative feedback, can determine how the student will interpret the verbal feedback. Burnett (2002) proposed that a student's relationship with their teacher was negatively compromised when administered frequent negative feedback. Furthermore, students provided consistent positive feedback on their performance in the classroom reported to have higher perceptions of a positive learning environment. This is not to say that teachers will only give positive feedback, but to be aware of the tone and connotation if the observation is negative. The tone of a teacher's verbal feedback is an additional consideration that is central to assessing the quality of verbal feedback.

Purpose

There is an extensive body of research exploring the quality of feedback that informs pre-service teachers regarding their teaching practices and the most effective teaching practices used by teachers in the field of music (Bernard, 2009; Legette & McCord, 2015; Richards & Killen, 1993; Walker, 2008; White, 2007, for example.). However, the quality of formative, verbal feedback has not been directly examined in the context of pre-service teachers in the field of music education. Therefore, a measurement instrument that can be used to engage pre-service music educators in developing effective verbal feedback processes in the secondary music classroom may offer an important and meaningful mechanism for both assessing and supporting teacher-student dialogue regarding this important aspect of music teaching. The purpose of this study was the development and validation of a scale to measure the construct of quality pre-service music educators' formative, verbal feedback in the context of secondary-level instrumental ensemble rehearsals. The Pre-Service Music Teacher Verbal Feedback Evaluation Scale was constructed for three reasons: (a) as a tool to help guide, facilitate, develop, and

engage pre-service music educators' in these considerations, (b) to provide guidelines and pedagogical talking points for the criteria encompassed by effective verbal feedback in music education, and (c) to empirically examine how the construct of quality of pre-service music educators' verbal, formative feedback manifests in order to facilitate better pre-service teacher instruction. The research questions that guided the study include:

1. What are the psychometric qualities (i.e., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale?
2. How do the verbal feedback criteria vary in difficulty?
3. How does the rating scale category structure vary across each individual criterion?

Method

Participants and Raters

Pre-service music educators ($N = 55$) from a large southeastern university in the United States were video recorded in 6-10-minute teaching segments. Teaching segments were randomly assigned an ID code from 1-55 in order to protect the confidentiality of the pre-service music educators in the teaching segments. All participants were pre-service, undergraduate music educators at the time the data were collected. Participants ranged between the academic years of undergraduate sophomores to seniors ($N = 55$, male $n = 36$, female $n = 19$). The pre-service teachers volunteered for the study with informed consent. A standard lesson plan template was provided to the pre-service teachers and explained to them in detail. Each pre-service teacher was asked to prepare a lesson plan using the template and to determine the central focus of their lesson. Pre-service teachers were aware their verbal feedback would be evaluated, but they were not provided details of the rating scale criteria before completing their lesson. This was done to ensure the authenticity of the individual's teaching segment. Pre-service teaching

was video recorded in a lab setting (i.e., an ensemble of their peers) during instrumental methods classes (i.e., woodwind methods, brass methods, and secondary methods courses). Pre-service teachers freely used materials of their choosing, such as method books, sheet music, a white board, an overhead projector, and their instrument for modeling during their teaching segment.

A total of 15 music content experts (i.e., raters) volunteered to participate in this study based on the criteria that they were an in-service music educator with experience supporting pre-service teachers through either practicum or student teaching experiences. Raters had an average of 14 years teaching experience (SD = 9.92). Video teaching segments were randomly assigned to raters in order to be evaluated using the proposed Pre-Service Music Teacher Verbal Feedback Evaluation Scale. The scale was disseminated to raters via a Google form, where they selected and submitted their responses for each of their assigned videos. Provided instructions explained how to navigate and use the form to evaluate the teaching segments based on the criteria of the rating scale. The raters were asked to watch the entire segment for each teacher they evaluated before filling out the rating scale, and a separate form was completed for each video. Raters were also provided a PDF of the scale prior to assignment of the teaching segments in order to familiarize themselves with the criteria and rating structure of the scale (Winter, 1993).

The rater assessment network was an incomplete balanced design (Linacre et al., 1994). Each rater was asked to evaluate six pre-service music educator teaching segments using the Pre-Service Music Teacher Verbal Feedback Evaluation Scale. Each consecutive rater evaluated three overlapping teaching segments in order to establish balanced connectivity (e.g., rater 1 evaluated teaching segments 1-6, rater 2 evaluated teaching segments 3-9, rater 3 evaluated teaching segments 6-12). This design was chosen in order to maximize raters' time while also providing strong support for data-to-model fit (Wesolowski, 2016; Wind et al., 2018). There was

overlap between each rater's observed teaching segments to allow for the variability and fit of rater responses to be assessed. Raters were provided an assigned rater number for anonymity purposes. In total, 84 observations were used in the analysis, providing a sufficient sample size to meet the requirements for stable and productive measurement (Wright & Stone, 1976; Linacre, 1994; Wright & Tennant, 1996; Smith et al., 2008; Bond & Fox, 2015).

Measurement Instrument

The domains and criteria for the scale were gleaned from the educational research literature on verbal feedback described above. Evidence of the content validity of domains and criteria was an essential part of the item construction and validation process (Kane, 2006). These criteria were chosen based upon their relevance to the performance-based setting of the instrumental music classroom and adapted into suitable criteria used to operationally define the construct: quality of pre-service music educators' administration of verbal feedback. In order to establish both content and face validity of the scale, the pool of criteria was screened and evaluated by the authors as well as three additional university professors specializing in secondary-level instrumental music education and pre-service music teacher preparation. The professors used the scale to preliminarily evaluate 30 pre-service music students and the resulting data were examined for outliers in students or criteria (Wright & Stone, 1976). Based on the results, adjustments to wording and anchor considerations were made. These students and professors' ratings were not included in the data for this study. Individual criteria were screened for clarity and ease of understanding, as well as cohesiveness. The scale included criteria (N = 39) grouped into five domains: (a) context/audience (n = 11); (b) learning objective/focus (n = 8); (c) timing (n = 5); (d) type (n = 11); and (e) tone (n = 4). The rating scale structure was based on a four-point Likert-type scale (see Figure 1.1). A four-point Likert scale response set

was chosen with no neutral response given for any item on the rating scale. The elimination of a neutral response allowed for a monotonic structure across the rating scale response set (Wright, 1977). Likert-type scale response anchors were selected based upon the underlying purpose of each criterion. Specifically, there were three sets of anchors used in the scale: (a) frequency (*never, rarely, sometimes, often*), (b) agreeability (*strongly disagree, disagree, agree, strongly agree*), and (c) appropriateness (*inappropriate, slightly inappropriate, slightly appropriate, appropriate*). A four-point, polytomous gradation of frequency, agreeability, and appropriateness anchors were chosen rather than a dichotomous response set (e.g., yes/no; agree/disagree) in order to eliminate the possibility of an acquiescence response bias or positivity bias, both of which may compromise the validity of the scale (Cronbach, 1942; Saris et al., 2010; Keep, 2019). The Pre-Service Music Teacher Verbal Feedback Evaluation Scale was specifically tested with the focus and intention of examining verbal feedback in the context of the secondary-level instrumental pre-service music educator, and any inferences gleaned through the validation process is only relevant in this specific context (Wright & Stone, 1999).

Psychometric Considerations

Item Response Theory (IRT) is a family of measurement models used to measure unobservable, latent constructs (Wesolowski, 2019). Rasch measurement models are a specific family of measurement models associated with IRT. In the event that raters mediate the assessment context, such as in this study, the Rasch family of measurement models is particularly useful due to their requirements of rater-mediated invariant measurement (Engelhard, 2013; Engelhard & Wind, 2018). For this study, the five requirements of rater-mediated invariant measurement can be interpreted as follows: (a) the measurement of students must be independent of the particular raters that happen to be used in the assessment (i.e., rater-

invariant measurement of students); (b) the calibration of the criteria must be independent of the particular raters used in the assessment (i.e., rater-invariant calibration of criteria); (c) the structure of the rating scale categories must be independent of the particular raters used in the assessment (i.e., rater-invariant calibration of rating scale categories); (d) the locations of raters must be independent of the particular students, criteria, and rating scale categories used in the assessment (i.e., invariant locations of raters); and (e) students, criteria, rating scale categories, and raters must be simultaneously located on an underlying latent continuum used in an assessment system (i.e., unidimensionality as evidenced by a rater-invariant Wright map). When adequate invariant measurement is actively obtained, sample-independent measures are achieved. More specifically, (a) student measures (i.e., student achievement) are estimated without being affected by the variability in criteria difficulty or the variability in rater severity of the particular sample; (b) criteria measures (i.e., criteria difficulty) are estimated without being affected by the variability in student achievement and variability in rater severity of the particular sample; and (c) rater measures (i.e., rater severity) are estimated without being affected by the variability in student achievement and variability of criteria difficulty of the particular sample.

The data were analyzed using the Many Facet Rasch Partial Credit (MFR-PC) model in order to examine the psychometric quality (i.e., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale. There are three important statistical indices examined in this analysis: (a) logit scale locations, (b) separation, and (c) data-to-model fit. Logit scale locations provide an empirical measure to examine the locations of the elements (i.e., each student, each item, and each rater) measured in this study. Each individual student, criterion, and rater is displayed on a Wright map (see requirement five of rater-mediated invariant measurement described above) in order to determine where they are located based on

the latent construct being measured. Separation statistics identify the significance of the spread of the students, criteria, and raters based upon their logit scale locations, and whether the students, criteria, and raters can be significantly differentiated. Data-to-model fit describes the behavior of the patterns of responses (Linacre, 2002) and to what degree invariant measurement is achieved (Engelhard, 2013).

The Many Facet Rasch Partial Credit Model

The Rasch measurement model can be used to simultaneously and independently estimate student achievement measures, criteria difficulty measures, and rater severity measures. The MFR-PC was adapted from the Many Facet Rasch (MFR) measurement model (Linacre, 1989/1994). In particular, the PC version of the model adds an additional interaction parameter that allows for the examination of the rating scale category thresholds for each individual criterion in the model (Masters, 1982). The rating scale category thresholds are an important empirical measure to investigate the precision points of the rating scale structure, as they inevitably vary for each individual criterion. The model specifications in this study included three facets: (a) students (i.e., student achievement), (b) raters (i.e., rater severity), and (c) criteria (i.e., difficulty of criteria). The FACETS (Linacre, 2014) computer program was used for the MFR-PC data analysis.

Results

Summary Statistics (Research Question 1)

Research question 1 addressed the psychometric qualities (i.e., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale. Table 1.1 provides the summary statistics for the MFR-PC model. The table provides statistics for three facets: (a) students (θ), (b) raters (λ), and (c) criteria (δ). The chi-square test of significance

answers the substantive question, "*Is there a statistical difference in the logit-scale locations for elements of each facet (i.e., all students, all raters, all criteria)?*" The chi-square test of significance for all students demonstrated a significant difference in students' overall logit-scale locations, $\chi^2_{(55)} = 986.30$, $p < .01$. The chi-square test of significance for all raters demonstrated a significant difference in raters' overall logit-scale locations, $\chi^2_{(15)} = 708.40$, $p < .01$. Lastly, the chi-square test of significance for all criteria demonstrated a significant difference in criteria's overall logit-scale locations, $\chi^2_{(39)} = 1000.10$, $p < .01$. The reliability of separation (*Rel*) statistic was examined in order to address the substantive question, "*Is there a significant spread of the different facets located on the logit-scale?*" There were high reliability measures for students (*Rel* = .94), criteria (*Rel* = .97), and raters (*Rel* = .98), indicating that the spread within each facet on the scale was significant. Model fit was examined to address the substantive question, "*Are the characteristic response patterns adequately consistent for the students, criteria, and raters?*" Item-mean square (MSE) indices indicate that students, criteria, and raters displayed adequate data-to-model fit using a fit indicator of 0.60-1.40 (see Table 1.1).

Criteria and Domain Calibrations (Research Question 2)

Research question 2 addressed the interpretation of the criteria and domains on the Pre-Service Music Teacher Verbal Feedback Evaluation Scale. The higher the logit score, the more difficult the criteria and/or domain. The lower the logit score, the less difficult the criteria and/or domain. The most difficult criterion (see Table 1.2) on the Pre-Service Music Teacher Verbal Feedback Evaluation Scale was Q5.03: *Teacher uses simple/understandable vocabulary when administering feedback* (-1.48 logits). The difficulty of this item indicates that pre-service music educators have difficulty providing simple and understandable feedback to their students. The least difficult criterion on the scale was Q2.08: *Teacher makes a connection to real life situations*

when giving feedback. (1.67 logits). The most difficult domain was learning objective/focus (see Table 1.2). The difficulty of this domain illustrates the difficulty pre-service music educators experience when trying to relate their feedback to the learning objective they are teaching throughout their lesson. The least difficult domain was tone. This illustrates that pre-service music educators often speak to their students using a respectful and positive tone. Criteria Q1.11, Q3.04, Q4.05, Q 4.11, and Q5.02 did not show a characteristic fit to the model, indicating unusual patterns in responses (see Table 1.2). It is therefore suggested that these items be either removed or edited based upon substantive considerations for future scale validation purposes (see Figure 1.2).

Structure of the Rating Scale (Research Question 3)

Research question 3 addressed how the rating scale category structure might vary across each individual criterion. The Many Facet Rasch Partial Credit (PC) Model (Masters, 1982) allows for the investigation of differing thresholds over separate domains and criteria of the scale. Therefore, the PC Model allows each criterion's rating scale structure to function as a separate entity. The rating scale structure for each criterion in this study is comprised of four response categories and the anchor content is specific to the appropriateness of the attribute being evaluated. Specifically, the anchors/categories used in the rating scale are (a) level of agreement; *Disagree/Somewhat Disagree/Somewhat Agree/Agree*, (b) level of appropriateness; *Inappropriate/Slightly Inappropriate/Slightly Appropriate/Appropriate*, and (c) frequency; *Never/Rarely/Sometimes/Often*. Rasch-Andrich thresholds provide category discrimination between neighboring categories at an interval level of measurement. Four categories are used for all criterion in the rating scale, allowing for three discrimination indices: (a) discrimination between Category 1 and Category 2, (b) discrimination between Category 2 and Category 3, and

(c) discrimination between Category 3 and Category 4. Evaluation of the Rasch-Andrich thresholds provides detailed information on the difficulty-level of moving between each adjacent rating scale category, confirming that the difficulties of each rating scale structure are uniquely attributed to the content of each criterion. The result provides more information and precision related to the ability of the students being evaluated.

Bond and Fox (2015) highlight the importance of analyzing the rating scale structure from a scale construction perspective. Rasch-Andrich thresholds allow for the post hoc examination of rating scale domains and criteria. Each time a scale is used, recalibration is important in order to evaluate scale construction and revision, as post hoc changes to the rating scale structure provide better precision to the future measurement process, and ultimately, a stronger argument for construct validity and item effectiveness. Table 1.3 displays category usage by frequency and percentage used, average observed logit measures and the average expected logit measure, and outfit mean squares (MSE). Linacre (2002) provides important considerations for optimizing the rating scale structure effectiveness. Any frequency count for any category which exhibits less than 10% usage provides grounds for the neighboring categories to be collapsed. Based upon the collected data, the following categories demonstrated less than 10% usage: Q1.02 category 1, Q1.03 category 1, Q1.04 category 1, Q1.05 categories 1 and 2, Q1.06 category 1, Q1.11 categories 3 and 4, Q2.08 category 4, Q3.01 category 1, Q3.02 categories 1 and 2, Q3.05 category 1, Q4.01 category 1, Q4.04 category 1, Q4.05 category 1, Q4.06 category 4, Q4.07 category 4, Q4.11 category 1, Q5.01 category 1, Q5.02 category 1, Q5.03 categories 1 and 2, and Q5.04 category 1. It is recommended that for future applications of the measurement instrument, these categories either be collapsed or if kept for substantive purposes, reexamined carefully. The average expected logit measure provides insights into the

monotonic structure of the rating scale categories, where it is expected to have an increasing average measure across all rating scale categories. Based upon the collected data, no violations of monotonicity occurred. Lastly, outfit mean squares (MSE) where values exceeding >/= 2.0 are cause for concern, as they imply unexpected response patterns. Based upon the collected data the following categories demonstrated MSE values >/= 2.0: Q1.01 category 2, Q1.03 category 1, Q1.11 categories 1, 2, 3, and 4, Q3.04 categories 1 and 4, and Q4.05 categories 1, 2 and 4. For future applications, it is recommended that these categories be collapsed or if kept for substitutive purposes, reexamined carefully.

Conclusion

The purpose of this study was to develop and validate a scale used to assess the quality of pre-service music educators' verbal feedback in the context of secondary-level instrumental ensemble rehearsals. This was accomplished by developing a scale that was psychometrically sound and could be used pedagogically as a tool for pre-service music educator assessment. The purpose of research question 1 (*What are the psychometric qualities (i.e., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale?*) was to determine whether this scale was a valid and reliable tool to evaluate pre-service teachers' verbal feedback. The results suggest that the Pre-Service Music Teacher Verbal Feedback Evaluation Scale, based upon the spread of the logit scale measures, separation statistics, and data-to-model fit indices, demonstrated strong validity, reliability, and precision in the measurement of pre-service music educators.

The purpose of research question 2 (*How do the verbal feedback criteria vary in difficulty in relation to how the students (i.e., pre-service teachers) vary in achievement?*) was to examine the ordering of criteria difficulty for pre-service teachers to master. The results of the second

research question suggest a clear ordering of domain- and criteria-level difficulty. The domain *Learning Objective/Focus* as a whole was the most difficult for teachers to achieve, followed by *Type, Context/Audience, Timing,* and the least difficult, *Tone.* Pedagogically, the understanding of difficulty ordering is a powerful instructional tool that may help facilitate important dialogue between pre-service instructors and pre-service teachers and provide a mechanism for curricular-ordering and topic discussion in pre-service teacher's coursework.

The results of the third research question (*How does the rating scale category structure vary across each individual criterion?*) suggest that the varying difficulties of the rating scale categories across each criterion are not equidistant and each criterion has a unique difficulty level across each of their respective rating scale category structures. Furthermore, some criteria had categories that were not used more than 10% of the time. A usage of less than 10% presents the argument for collapsing the neighboring categories as there are not particularly meaningful for evaluating the pre-service music teachers. The results from this study indicate that the Pre-Service Music Teacher Verbal Feedback Evaluation Scale is valid, reliable, and can be utilized as a powerful teaching tool for university professors training pre-service music teachers.

Discussion

Pre-service and early career music educators often feel under-prepared and under-developed as teachers when they begin their careers (Ballantyne, 2007). Berg and Miksza (2010) suggest that young teachers provide few instructions, little significant feedback, and little opportunity for students to apply any feedback given during their instruction. There is a reported disconnect between what students are taught in their pre-service programs and what they utilize as young music educators in the classroom. In a survey administered by Book et al. (1983), results showed that pre-service teachers placed a higher value on their field experience training

than their studies and preparation courses in music education (e.g., methods courses), highlighting the belief that the authors refer to as "experience is the best teacher." As they noted, "It is disturbing that preservice teachers by and large do not perceive a strong need to obtain a knowledge base in pedagogy in order to become effective teachers" (p. 10-11). In a later study by Richards and Killen (1993) it was suggested that, during their practicum teaching (e.g., field teaching while enrolled in undergraduate education), pre-service teachers tended to disregard theoretical and pedagogical knowledge learned during their undergraduate studies. Due to the feelings of unpreparedness exhibited by music educators entering the field as well as pre-service teachers' dispositions and assumptions exhibited by the "I'll know it when I see it" approach to understanding and teaching in music education, there is a need for the creation of meaningful mechanisms to teach, engage, and assess the quality of important teaching skills throughout teacher preparation programs (Parkes et al., 2019). These frameworks may help pre-service teachers realize what skills they lack so they can remedy them and make necessary and permanent changes to their teaching habits before entering the classroom.

Pre-service teachers' perceptions of what effective teaching looks like can be a hinderance on the learning and implementation of new knowledge and theoretical concepts taught in teacher training programs (Butler, 2001). In a synthesis of the literature, Berliner (2001) concluded that experienced and effective classroom teachers, as opposed to novice teachers, are more easily able to adapt to their students' instructional needs in the moment, based on their formative assessment. Going forward, it is important for pre-service educators to develop these skills, in order to promote student-centered learning and to accurately and effectively demonstrate student achievement in the classroom (Book et al., 1983).

Santagata and Angelici (2010) performed a study where pre-service teachers observed teaching videos of other educators. They were then prompted to analyze and evaluate these videos as part of their university course. In turn, the analysis of the teaching videos facilitated more frequent self-reflection on their own teaching and practice (Santagata & Angelici, 2010). Similar to the evaluation framework used in the study by Santagata and Angelici (2010), the Pre-Service Music Teacher Verbal Feedback Evaluation Scale is an advantageous teaching tool which can help pre-service music educators, in collaboration with their university professors, to evaluate their teaching and feedback quality to see how their skills improve and grow throughout their pre-service teaching career. Future studies might include performing an accuracy model (e.g., comparative study) between professor and pre-service music educator, where the professor and pre-service music educator would evaluate the student using the Pre-Service Music Teacher Verbal Feedback Evaluation Scale. This would allow researchers to compare the way pre-service music educators perceive the quality of their verbal feedback versus the way their professor views the same feedback. Comparing these two perceptions could help pre-service music educators grow in the way where they perceive the quality of their own verbal feedback, and in turn, help them to become more effective teachers.

Table 1.1

Summary Statistics from the Many Facet Rasch Partial Credit Model

	Students (θ)	Criteria (λ)	Raters (δ)
Logit-Scale Location			
M	0.34	0.00	0.00
SD	0.76	0.82	0.68
N	55	39	15
Infit MSE			
M	1.03	1.01	1.03
SD	0.33	0.56	0.27
Std. Infit			
M	0.00	-0.40	-0.10
SD	1.70	2.60	2.90
Outfit MSE			
M	1.07	1.09	1.10
SD	0.38	0.92	0.30
Std. Outfit			
M	0.20	-0.50	0.50
SD	1.70	2.70	2.80
Separation Statistics			
Reliability of Separation	0.94	0.97	0.98
Chi-Square	986.30*	1000.10*	708.40*
Degrees of Freedom	54	38	14

Note. *$p < .01$.

Table 1.2

Calibration of Domain and Criteria Facets

	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
Domain							
Learning Objective/Focus	2.45	0.45	0.14	0.84	-1.20	0.81	-1.30
Type	2.60	0.27	0.15	1.09	0.00	1.14	-0.10
Context/Audience	2.80	-0.17	0.15	1.00	-0.70	1.17	-0.70
Timing	2.91	-0.24	0.16	1.27	1.10	1.51	1.00
Tone	3.17	-0.91	0.17	0.85	-1.10	0.83	-1.30
Criteria							
Q1.11	1.39	2.02	0.17	2.76	5.00	4.57	4.40
Q2.08	1.63	1.67	0.16	0.95	-0.20	0.84	-0.60
Q4.07	2.06	1.21	0.15	1.22	1.50	1.21	1.30
Q4.08	2.00	1.06	0.14	1.09	0.60	1.12	0.60
Q3.04	2.04	1.04	0.14	2.72	8.00	4.16	9.00
Q4.06	2.17	0.92	0.15	1.02	0.20	0.98	-0.10
Q2.07	2.19	0.88	0.14	0.88	-0.80	0.85	-1.00
Q4.03	2.17	0.87	0.15	1.00	0.00	1.03	0.20
Q2.05	2.35	0.55	0.13	0.82	-1.20	0.78	-1.30
Q4.02	2.40	0.49	0.14	1.05	0.40	1.11	0.70
Q1.07	2.48	0.44	0.14	0.73	-2.10	0.74	-1.90
Q2.04	2.52	0.36	0.14	0.84	-1.10	0.82	-1.20
Q3.03	2.66	0.18	0.13	0.93	-0.50	0.87	-0.80
Q4.04	3.07	0.18	0.17	0.66	-3.00	0.63	-3.00
Q2.01	2.63	0.16	0.14	0.96	-0.20	0.94	-0.30
Q1.01	2.67	0.13	0.13	1.16	1.10	1.24	1.40
Q4.09	2.67	0.11	0.14	0.79	-1.60	0.75	-1.80
Q1.10	2.73	0.06	0.15	0.65	-2.70	0.65	-2.60
Q2.03	2.70	0.04	0.14	0.63	-2.90	0.62	-2.90
Q1.08	2.71	0.00	0.15	0.68	-2.40	0.69	-2.40
Q2.06	2.78	-0.01	0.14	0.71	-2.10	0.70	-2.10
Q2.02	2.77	-0.03	0.15	0.91	-0.60	0.92	-0.40
Q4.10	2.80	-0.06	0.14	0.67	-2.50	0.65	-2.60
Q1.09	2.77	-0.08	0.14	0.83	-1.20	0.81	-1.20
Q4.05	2.76	-0.22	0.15	3.13	9.00	3.91	9.00
Q4.11	2.90	-0.33	0.15	0.58	-3.30	0.57	-3.30
Q1.06	2.94	-0.39	0.15	0.71	-2.10	0.70	-2.10
Q1.02	3.08	-0.50	0.14	0.85	-0.90	0.94	-0.30

Q5.04	3.02	-0.50	0.15	0.93	-0.30	0.90	-0.60
Q3.01	3.34	-0.58	0.19	0.90	-0.70	0.83	-1.10
Q3.05	3.13	-0.66	0.15	0.94	-0.30	0.95	-0.20
Q5.01	3.13	-0.79	0.16	0.88	-0.70	0.82	-1.10
Q5.02	3.08	-0.87	0.16	0.55	-3.60	0.54	-3.60
Q1.04	3.23	-1.06	0.16	0.87	-0.80	0.84	-0.80
Q1.03	3.45	-1.14	0.16	0.92	-0.40	0.89	-0.30
Q3.02	3.36	-1.17	0.17	0.84	-0.80	0.73	-1.50
Q4.01	3.64	-1.21	0.20	0.73	-1.50	0.56	-1.70
Q1.05	3.31	-1.33	0.19	0.80	-1.20	0.76	-1.50
Q5.03	3.43	-1.48	0.18	1.05	0.30	1.07	0.40

Note. The criteria are arranged from high to low (e.g., most difficult to least difficult).

Table 1.3

Category Diagnostics: Category Usage, Average Observed/Expected Measures, and Outfit MSE

Criteria	Category Usage (%)				Average Observed Logit Measure (Average Expected Logit Measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
1.01	17(20)	17(20)	25(30)	24(29)	-0.38(-0.62)	-0.01(-0.16)	-0.14(0.35)	1.21(0.99)	1.4	†2.2	1.2	0.6
1.02	§7(8)	12(14)	31(37)	33(40)	-0.25(-0.23)	0.27(0.23)	0.57(0.75)	1.58(1.42)	1.1	1.2	0.7	0.8
1.03	§3(4)	9(11)	19(23)	52(63)	1.15(0.22)	0.10(0.67)	1.09(1.18)	1.92(1.84)	†2.9	0.2	0.5	0.9
1.04	§2(2)	16(19)	26(31)	39(47)	-0.59(0.22)	0.68(0.70)	1.24(1.23)	1.95(1.91)	0.5	0.8	0.8	1.0
1.05	§1(1)	§7(8)	40(48)	35(42)	-0.07(0.36)	0.51(0.86)	1.40(1.45)	2.31(2.18)	0.6	0.6	0.7	0.9
1.06	§6(7)	19(23)	32(39)	26(31)	-0.43(-0.29)	-0.05(0.19)	0.74(0.74)	1.62(1.42)	1.0	0.4	0.6	0.8
1.07	15(18)	25(30)	31(37)	12(14)	-1.04(-0.90)	-0.49(-0.39)	0.21(0.19)	1.24(0.90)	0.8	0.5	0.9	0.7
1.08	8(10)	24(29)	35(42)	16(19)	-0.67(-0.59)	-0.40(-0.08)	0.63(0.50)	1.43(1.21)	0.9	0.4	0.6	0.8
1.09	11(13)	22(27)	25(30)	25(30)	-0.56(-0.48)	-0.03(-0.01)	0.39(0.51)	1.35(0.51)	0.8	0.8	1.1	0.7
1.10	10(12)	18(22)	39(47)	16(19)	-0.99(-0.65)	-0.23(-0.16)	0.37(0.41)	1.54(1.13)	0.6	0.6	0.8	0.7
1.11	65(78)	8(10)	§6(7)	§4(5)	-1.59(-1.91)	-2.14(-1.25)	-1.90(-0.58)	-1.49(-0.03)	†2.8	†3.4	†5.0	†7.1
2.01	11(13)	25(30)	31(37)	16(19)	-0.56(-0.69)	-0.29(-0.18)	0.36(0.38)	1.21(1.08)	1.1	0.8	0.9	0.8
2.02	8(10)	19(23)	40(48)	16(19)	-0.47(-0.59)	-0.18(-0.09)	0.39(0.49)	1.52(1.22)	1.2	0.8	0.8	0.8
2.03	10(12)	23(28)	32(39)	18(22)	-0.73(-0.61)	-0.35(-0.11)	0.44(0.45)	1.54(1.15)	0.8	0.6	0.4	0.6
2.04	16(19)	23(28)	29(35)	15(18)	-0.83(-0.82)	-0.42(-0.32)	0.17(0.23)	1.21(0.92)	1.1	0.7	0.7	0.7
2.05	26(31)	20(24)	19(23)	18(22)	-0.95(-0.88)	-0.36(-0.40)	-0.06(0.12)	0.99(0.74)	0.9	0.8	0.9	0.6
2.06	11(13)	16(19)	36(43)	20(24)	-0.77(-0.58)	-0.15(-0.10)	0.33(0.45)	1.50(1.14)	0.7	0.9	0.6	0.7
2.07	25(30)	26(31)	23(28)	9(11)	-1.13(-1.18)	-0.81(-0.66)	-0.09(-0.07)	0.97(0.62)	1.0	0.9	0.8	0.6
2.08	49(59)	21(25)	8(10)	§5(6)	-1.75(-1.70)	-1.04(-1.09)	0.00(-0.43)	-0.16(0.18)	0.9	0.6	0.7	1.2
3.01	-	9(11)	37(45)	37(45)	-	-0.06(0.10)	0.63(0.68)	1.51(1.41)	-	0.8	0.6	1.0
3.02	§2(2)	§7(8)	33(40)	41(49)	-0.48(0.22)	0.70(0.70)	1.18(1.26)	2.05(1.96)	0.4	0.8	0.5	1.0
3.03	16(19)	16(19)	31(37)	20(24)	-0.47(-0.68)	-0.50(-0.21)	0.12(0.31)	1.34(0.98)	1.3	0.5	0.8	0.6
3.04	31(37)	27(33)	16(19)	9(11)	-0.24(-1.26)	-1.05(-0.72)	-0.70(-0.13)	-0.99(0.53)	†4.1	1.3	1.7	†8.7
3.05	§5(6)	14(17)	29(35)	35(42)	0.01(-0.11)	0.31(0.36)	0.79(0.89)	1.64(1.56)	1.5	0.9	0.6	0.9
4.01	-	6(7)	18(22)	59(71)	-	0.30(0.56)	0.79(1.12)	1.95(1.82)	-	0.6	0.4	0.8
4.02	15(18)	32(39)	24(29)	12(14)	-0.82(-0.91)	-0.42(-0.39)	0.22(0.20)	0.82(0.89)	1.1	0.7	1.4	1.2
4.03	20(24)	37(45)	18(22)	8(10)	-0.94(-1.19)	-0.79(-0.64)	-0.12(-0.01)	1.02(0.69)	1.2	1.0	1.2	0.7

4.04	-	21(25)	35(42)	27(33)	-	-0.68(-0.48)	-0.05(0.10)	1.18(0.82)	-	0.7	0.4	0.6
4.05	§5(6)	28(34)	32(39)	18(22)	2.17(-0.41)	1.31(0.11)	0.15(0.70)	-0.20(1.41)	†6.5	†4.2	1.3	†3.5
4.06	20(24)	36(43)	20(24)	§7(8)	-1.20(-1.24)	-0.75(-0.69)	0.09(0.48)	0.48(0.66)	1.1	0.8	0.9	1.1
4.07	29(35)	25(30)	24(29)	§5(6)	-1.26(-1.46)	-1.03(-0.92)	-0.44(-0.29)	0.52(0.41)	1.4	1.1	1.2	0.9
4.08	36(43)	21(25)	16(19)	10(12)	-1.09(-1.25)	-0.88(-0.73)	-0.55(-0.16)	0.89(0.48)	1.3	1.1	1.5	0.5
4.09	13(16)	21(25)	29(35)	20(24)	-0.75(-0.63)	-0.19(-0.15)	0.30(0.39)	1.32(1.06)	0.8	0.7	0.8	0.7
4.10	11(13)	18(22)	31(37)	23(28)	-0.86(-0.52)	0.04(-0.05)	0.31(0.49)	1.50(1.16)	0.6	0.8	0.5	0.7
4.11	§6(7)	19(23)	35(42)	23(28)	-0.85(-0.35)	-0.03(0.14)	0.66(0.70)	1.73(1.40)	0.6	0.5	0.4	0.7
5.01	§3(4)	13(16)	37(45)	30(36)	-0.19(-0.03)	0.14(0.46)	1.15(1.02)	1.73(1.73)	0.9	0.4	0.7	1.1
5.02	§2(2)	17(20)	36(43)	28(34)	-0.85(0.06)	0.31(0.57)	1.04(1.14)	2.20(1.85)	0.5	0.6	0.3	0.6
5.03	§1(1)	§7(8)	30(36)	45(54)	0.08(0.48)	1.45(0.96)	1.38(1.52)	2.25(2.22)	0.6	1.7	0.7	1.0
5.04	§5(6)	14(17)	38(46)	26(31)	-0.53(-0.24)	0.36(0.24)	0.75(0.80)	1.57(1.51)	0.6	1.1	0.8	1.0

Note. Category 1 = “Disagree/Inappropriate/Never;” Category 2 = “Somewhat Disagree/Slightly Inappropriate/Rarely;” Category 3 = “Somewhat Agree/Slightly Appropriate/Sometimes;” Category 4 = “Agree/Appropriate/Often.”

Criteria: See *Figure 1.1*.

§ Indicates category usage under 10%;

† Indicates outfit MSE >/= 2.00.

Table 1.4

Calibration of Student Facet

Teaching Segment	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
22	3.53	2.44	0.18	1.47	2.30	1.17	0.80
15	3.69	2.01	0.31	2.03	2.60	1.44	1.10
17	3.27	1.38	0.17	0.88	-0.70	0.80	-1.00
55	2.85	1.30	0.21	0.78	-1.00	0.75	-1.10
54	3.14	1.26	0.13	0.93	-0.50	0.87	-0.90
21	2.96	1.24	0.15	1.25	1.50	1.53	2.80
24	2.92	1.20	0.21	0.82	-0.80	0.90	-0.30
52	3.28	1.13	0.23	0.65	-1.60	0.65	-1.40
29	3.13	1.10	0.18	1.61	2.90	1.54	1.90
23	2.85	1.06	0.21	0.72	-1.30	0.79	-0.80
51	3.18	0.92	0.23	0.94	-0.10	0.83	-0.60
34	2.74	0.89	0.15	1.16	1.00	1.15	0.80
44	2.69	0.81	0.21	0.39	-3.70	0.42	-3.20
16	3.23	0.81	0.23	1.26	1.10	1.22	0.80
28	3.69	0.76	0.31	1.58	1.60	1.02	0.10
45	2.78	0.75	0.15	0.93	-0.40	0.89	-0.60
36	3.05	0.75	0.22	1.18	0.80	1.10	0.40
14	2.97	0.70	0.15	0.87	-0.80	0.87	-0.70
10	2.99	0.63	0.15	1.35	2.00	1.50	2.50
18	2.87	0.61	0.15	1.26	1.60	1.36	2.00
30	2.91	0.59	0.17	1.11	0.60	1.06	0.30
27	3.64	0.58	0.29	1.67	2.00	1.39	1.10
48	2.79	0.56	0.21	0.72	-1.30	0.67	-1.50
20	2.56	0.54	0.20	0.92	-0.30	0.95	-0.10
9	2.94	0.53	0.15	0.87	-0.80	1.02	0.10
4	3.03	0.49	0.22	1.06	0.30	1.06	0.30
32	2.08	0.42	0.21	0.97	0.00	0.96	0.00
49	2.77	0.35	0.15	0.72	-2.00	0.74	-1.60
39	3.03	0.33	0.22	1.36	1.40	1.17	0.70
31	2.03	0.33	0.21	1.27	1.20	1.27	0.90
26	2.97	0.31	0.16	1.11	0.60	1.02	0.10
50	2.73	0.29	0.15	0.63	-2.80	0.60	-2.80
11	2.51	0.28	0.20	0.93	-0.20	1.53	2.00
7	3.05	0.23	0.22	1.24	1.00	1.06	0.30
8	3.05	0.23	0.22	0.82	-0.70	0.83	-0.60
13	2.64	0.12	0.15	1.25	1.60	1.57	2.90
41	2.59	0.12	0.15	0.66	-2.50	0.65	-2.20
19	2.31	0.12	0.21	1.27	1.20	2.19	3.60

46	2.37	0.07	0.15	0.51	-4.10	0.51	-3.40
47	2.49	0.05	0.20	0.73	-1.30	0.69	-1.40
6	2.82	-0.03	0.15	0.96	-0.20	0.91	-0.50
42	2.49	-0.05	0.15	0.54	-3.70	0.51	-3.30
12	2.26	-0.14	0.21	0.53	-2.70	0.89	-0.30
5	2.62	-0.37	0.15	1.17	1.10	1.62	3.20
3	2.46	-0.47	0.20	1.00	0.00	1.10	0.50
2	2.44	-0.51	0.20	0.86	-0.60	0.97	0.00
43	1.85	-0.64	0.23	0.66	-1.60	0.85	-0.30
35	2.23	-0.64	0.21	0.85	-0.70	0.86	-0.40
33	1.86	-0.73	0.16	1.62	3.20	2.22	3.70
53	2.00	-0.74	0.13	1.06	0.40	1.48	2.40
25	2.40	-0.83	0.16	1.46	2.50	1.70	2.80
38	2.18	-0.91	0.15	0.95	-0.30	0.99	0.00
1	2.15	-.098	0.21	1.32	1.40	1.47	1.60
40	2.18	-1.09	0.21	0.68	-1.60	0.65	-1.40
37	1.97	-1.28	0.15	0.97	-0.10	1.14	0.70

Note. Students are arranged from high to low (e.g., highest achieving to lowest achieving).

Table 1.5

Calibration of Rater Facet

Rater	Observed Average Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
8	2.11	1.18	0.09	1.13	1.50	1.39	2.90
15	2.44	0.79	0.13	1.20	1.40	1.55	2.90
6	2.71	0.57	0.09	0.98	-0.10	1.14	1.20
11	2.32	0.55	0.09	0.50	-7.20	0.54	-5.20
5	2.81	0.52	0.09	1.26	2.70	1.48	4.10
3	2.56	0.32	0.08	0.92	-0.90	1.33	3.10
12	2.66	0.13	0.08	0.79	-2.50	0.77	-2.60
14	2.65	0.00	0.16	0.90	-0.60	0.90	-0.50
9	2.44	-0.16	0.09	1.25	2.60	1.28	2.40
13	2.94	-0.20	0.09	0.69	-3.80	0.67	-3.60
1	2.59	-0.39	0.08	1.02	0.30	1.25	2.40
4	3.25	-0.41	0.10	1.23	2.10	1.11	1.00
10	2.57	-0.53	0.09	0.81	-2.30	0.75	-2.70
2	3.05	-0.68	0.09	1.12	1.20	1.02	0.20
7	3.55	-1.68	0.11	1.63	4.60	1.34	2.10

Note. Raters are arranged from high to low (e.g., most severe to least severe).

Figure 1.1

Pre-Service Music Teacher Verbal Feedback Evaluation Scale

Domain	Rating Scale Categories			
Context/Audience				
1.01. Teacher provides individual feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.02. Teacher provides feedback to small groups of students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.03. Teacher provides feedback to the entire ensemble.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.04. Teacher chooses the appropriate method of feedback (e.g., individual or group) for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.05. The chosen type of feedback (e.g., oral, demonstration, nonverbal) is appropriate for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.06. Teacher addresses student errors during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.07. Teacher addresses student misconceptions during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.08. Teacher checks for student understanding of feedback.	Never	Rarely	Sometimes	Often
1.09. Teacher appropriately fields questions/responses from their students about feedback given.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.10. Teacher provides enough feedback for students to understand the objective.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.11. Teacher provides an overwhelming amount of feedback to students (e.g., too many learning objectives for the student to focus on at once).	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Learning Objective/Focus				

2.01. Teacher either states, or makes clear through their feedback, what the learning objective is for the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.02. The feedback administered focuses on the goal (e.g., teaching focus) of the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.03. Teacher provides feedback about the learning process.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.04. Teacher compares (by administering feedback) student work to established criteria.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.05. Teacher compares (by administering feedback) student work to past performances.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.06. Teacher provides steps for improvement when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.07. Teacher makes a connection to students' prior knowledge when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.08. Teacher makes a connection to real life situations when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Timing				
3.01. The timing of teacher feedback is appropriate for the event being addressed (e.g., it is given while students are still mindful of the learning target).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.02. Feedback is provided at an appropriate time (e.g., feedback relates to the learning task at hand).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.03. Teacher takes advantage of teachable moments to give feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
3.04. Administration of teacher feedback is delayed (e.g., event is not immediately addressed).	Disagree	Somewhat Disagree	Somewhat Agree	Agree
3.05. Students are presented opportunities to apply feedback once it is given.	Never	Rarely	Sometimes	Often

Type				
4.01. Teacher provides oral feedback.	Never	Rarely	Sometimes	Often
4.02 Teacher provides feedback by modeling/demonstrating.	Never	Rarely	Sometimes	Often
4.03. Teacher uses nonverbal feedback (e.g., picture, diagrams, gestures).	Never	Rarely	Sometimes	Often
4.04. Teacher chooses appropriate feedback content (e.g., relates to the lesson, or student/ensemble error).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
4.05. Teacher provides ambiguous/undescriptive feedback.	Never	Rarely	Sometimes	Often
4.06. Teacher uses feedback to prompt student discussion/reflection.	Never	Rarely	Sometimes	Often
4.07. Teacher compares the performance of a student/group of students to another student/group of students (norm-referenced).	Never	Rarely	Sometimes	Often
4.08. Teacher compares the performance of a student/group of students to a set of standards (criterion-referenced).	Never	Rarely	Sometimes	Often
4.09. Teacher uses self-referenced feedback (e.g., directly compares student performance to their previous performances).	Never	Rarely	Sometimes	Often
4.10. The function of teacher feedback is descriptive in nature.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
4.11. The function of teacher feedback is evaluative in nature.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Tone				
5.01. Teacher provides positive comments on student performance.	Never	Rarely	Sometimes	Often
5.02. Teacher provides constructive comments on student performance.	Never	Rarely	Sometimes	Often
5.03. Teacher uses simple/understandable vocabulary when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
5.04. Teacher is clear when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree

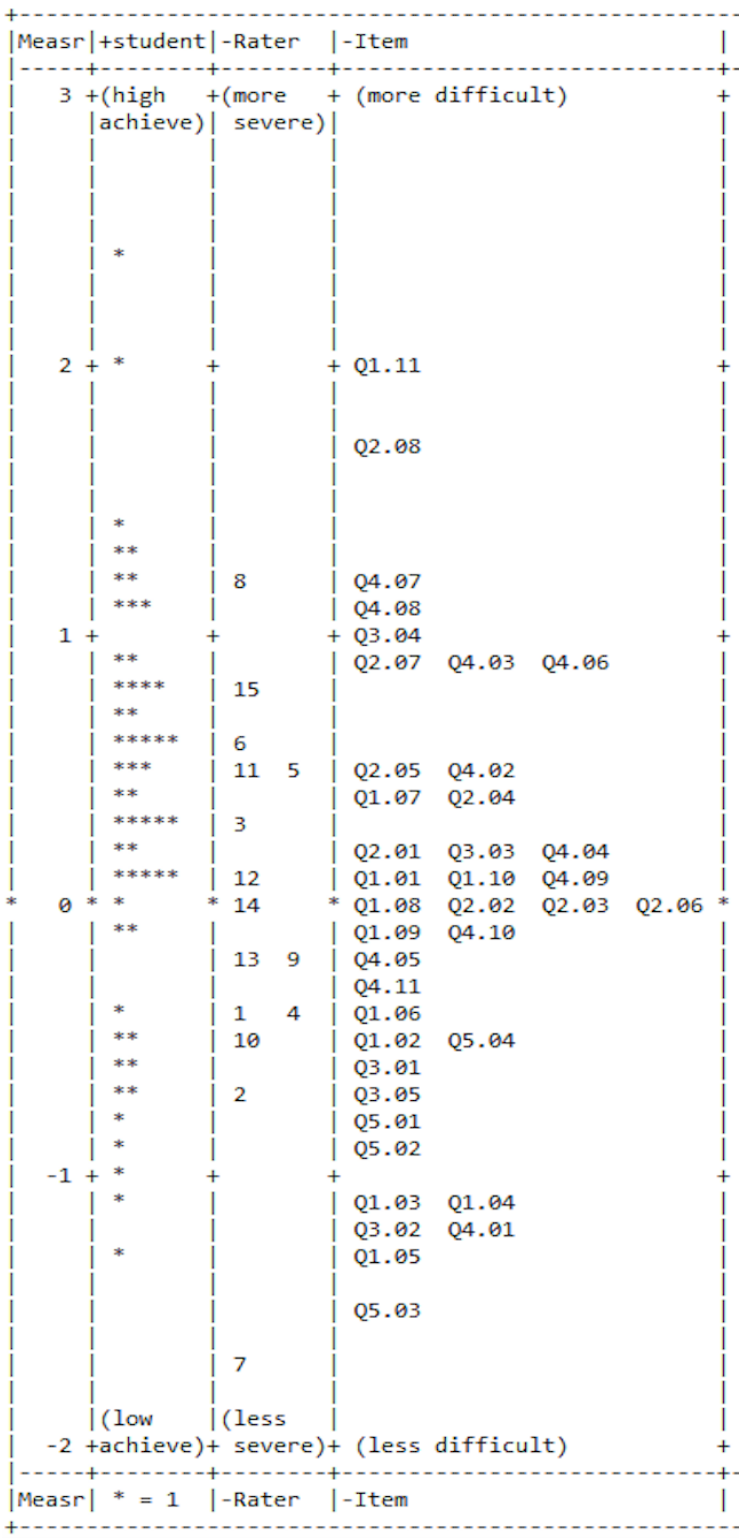
Figure 1.2

Scale Items to be Removed or Reexamined

Domain	Rating Scale Categories			
Context/Audience				
1.11. Teacher provides an overwhelming amount of feedback to students (e.g., too many learning objectives for the student to focus on at once).	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Timing				
3.04. Administration of teacher feedback is delayed (e.g., event is not immediately addressed).	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Type				
4.05. Teacher provides ambiguous/undescriptive feedback.	Never	Rarely	Sometimes	Often
4.11. The function of teacher feedback is evaluative in nature.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Tone				
5.02. Teacher provides constructive comments on student performance.	Never	Rarely	Sometimes	Often

Figure 1.3

Wright Map for the Many Facet Rasch Partial Credit Model



CHAPTER 3

EXAMINING SELF-ASSESSMENT ACCURACY OF PRE-SERVICE MUSIC EDUCATORS' QUALITY OF VERBAL FEEDBACK[2]

[2] Athanas, M.I. and B.C. Wesolowski. To be submitted to *Journal of Research in Music Education.*

Abstract

The purpose of this study was to examine the self-assessment accuracy of pre-service music educators' quality of verbal feedback in the context of secondary-level instrumental ensemble instruction. The questions that guided this study include: (a) Overall, how accurate were pre-service music educators' perceptions of their verbal feedback when compared to content experts' perceptions? (b) How does accuracy vary across each item of the scale? and (c) How does accuracy vary across each domain of the scale? Using a 35-item rating scale embedded within five instructional domains, the accuracy between content experts' evaluation of students and students' self-evaluation for the same teaching episode were examined at both item- and domain- levels. Considerations for the inclusion of self-assessment accuracy measures in teacher preparation curricula and its role in improving student-teacher communication, instructional quality, differentiated instruction, and reflective practice will be discussed.

Keywords: self-assessment, pre-service, verbal feedback, accuracy, Rasch

Introduction

In order for pre-service educators to effectively engage in and learn from self-assessment strategies throughout teaching and learning cycles, they must be part of an educational environment that cultivates self-regulated learning, accountability, and growth mindset (Carless, 2006; Nicol & Macfarlane-Dick, 2006). Regrettably, pre-service educators are not consistently and explicitly trained to reflect on their pedagogical practices through the use of methodical, research-based self-assessment procedures (Conway & Hibbard, 2020). Reflection on "real world" experiences with students is typically removed from any authentic teaching situation and transferred into the traditional university classroom setting (e.g., writing a reflection or a having a class discussion) which creates a divide in pre-service teacher understanding of how to properly assess the quality of their teaching while in the field (Rosaen & Florio-Ruane, 2008). In order for pre-service educators to be self-regulated learners focused on their personal growth through self-assessment, course curricula must provide space for authentic teaching experiences that promote purposeful and meaningful evaluations, along with the development of self-monitoring skills (Sadler, 1989). Developing a deep understanding of the methodologies used in their teacher education programs allows pre-service educators to improve their reflexive practices centered on their own pedagogical knowledge and growth (Saliba & Barrett, 1993).

A challenge facing the field of music education is the development of *authentic*, *purposeful*, and *high-quality* assessments that are "meaningful, measurable, and manageable" and grounded in theories of educational measurement (Kimpton and Kimpton, 2019). It is crucial to the teacher preparation process that teacher educators develop valid and reliable standards-based music assessments (e.g., grounded in educational measurement) with clear criterion for the primary purpose of improving teaching quality and student learning outcomes (Burrack and

Parkes, 2019). Strides are being taken in the field to develop standards-based assessments for P-12 students (the Model Cornerstone Assessments, for example) which align classroom assessment with the 2014 National Music Standards (Burrack and Parkes, 2018). However, the same strides are not being made to develop assessments to aid pre-service teachers in their preparation to enter the field as well-rounded music educators, based on the criteria deemed important in teacher education curriculums. In recent years, music teacher education programs in numerous states have been required to administer the Educative Teacher Performance Assessment (edTPA) to pre-service educators during their student teaching semester. The edTPA is a high-stakes summative assessment intended to measure teacher readiness for certification and graduation, and operates under the assumption that there is a significant correlation between certification and student preparedness to enter the classroom, and furthermore, that the edTPA accurately evaluates teacher quality or effectiveness (Darling-Hammond et al., 2000; Floden, 2008). Music teacher educators have founded validity and reliability concerns regarding the edTPA, such as the fact that the K-12 Performing Arts edTPA examines the arts as a whole instead of distinguishing by individual content area (i.e., music, dance, drama, and theater), raters' lack of contextual understanding of the teaching situation of the pre-service teacher they are evaluating, and the lack of feedback and justification for the score earned by the pre-service teacher (Musselwhite & Wesolowski, 2021). Musselwhite and Wesolowski (2021) found that, when examining the rubrics of the edTPA, pre-service music educators had difficulty with methods of assessment (i.e., formative and summative) and planning. They recommend pre-service teachers utilize formative assessment procedures to examine their students' knowledge in order to reflect upon and revise teaching practices to improve student learning outcomes (Musselwhite & Wesolowski, 2021). Due to the lack of a quality "standardized" pre-service

teacher assessment in music education which authentically examines music teaching, Parkes (2020) indicates a need for the construction and implementation of empirical formative assessments throughout the field experience and student teaching processes, as it is currently the pinnacle learning experience of the pre-service teacher training program, and quality learning cannot be achieved if goals are not articulated and presented in an authentic, yet decisive manner.

Reflective Practices in Teacher Education

Self-reflective practices in the field of education date back to the American philosopher John Dewey, who emphasized that self-improvement is derived from a reflective thought process, and therefore teaching cannot be detached from reflection. Dewey (1910) opined that self-reflection is a habit which requires a person to have mental discipline and to train the mind. He affirmed that:

> While it is not the business of education to prove every statement made…it is its business to cultivate deep-seated and effective habits of discriminating tested beliefs from mere assertions, guesses, and opinions; to develop a lively, sincere, and open-minded preference for conclusions that are properly grounded, and to ingrain into the individual's working habits, methods of inquiry and reasoning appropriate to the various problems that present themselves. (pp. 27-28)

Thus, no amount of experience can prepare pre-service music educators to be effective teachers if they are not taught to be proficient in the skill of self-assessment and reflection during their teacher training programs (Dinkelman, 2003). However, it has been repeatedly stated throughout the literature that *context matters,* and it is important for pre-service teachers to be trained in and to be given opportunities to reflect upon authentic experiences in similar contexts to which they

will teach (Brophy, 2000; Conway 2002, 2012). Educators are responsible for analyzing and reflecting on their own teaching practices and must do so accurately and efficiently in order to adjust their teaching plans and strategies to help students reach their goals (Darling-Hammond, 1999). The teacher education research community carries the responsibility of establishing and implementing reliable tools to equip future educators with high-quality measures for self-assessment (Howard & Aleman, 2008).

Students within higher education music degree programs often rely heavily on the knowledge and assessment of their professors instead of taking an active role in the self-assessment of their performances (Daniel, 2001). Due to the nature of the teaching profession, specifically in the field of music, teachers often find themselves isolated and lacking the opportunity to receive feedback (Freiberg, 1987). Therefore, in order to promote quality teaching methods and reflective practices beyond the teacher education program, pre-service teachers must develop self-assessment expertise in order to evaluate their own teaching effectiveness in the absence of a content expert (Sadler, 2010). Simple memory recall of teaching episodes is not a sufficient approach to accurate self-assessment (Gelfuso & Dennis, 2014). A more appropriate illustration of teaching would be the use of what Anderson and Freiberg (1995) call “living data” (i.e., video or audio recordings), where the teacher can watch their lesson and assess their teaching strategies, hence making video recordings a reliable resource to provide educators a medium for self-assessment, reflection, and personal growth that does not require the feedback of another party.

High-Quality Self-Assessment Tools for Teacher Training

In order for self-assessment to be effective, it is important to determine how accurately pre-service teachers evaluate their performances. Falchikov and Boud (1989) performed a meta-

analysis to investigate the results of quantitative research studies in higher education in order to evaluate whether students enrolled in college courses were able to accurately assess their performance achievement in comparison to their professors (e.g., content experts). Of particular interest was whether the students' self-assessment accuracy could improve with more experience and practice. Self-assessment measures with strong construct validity, such as a rating scale with specified criteria, rendered more accurate self-assessment results for participants. The more collegiate experience the participant had the more likely they were to achieve self-assessment accuracy, and pre-service teacher participants were noted to be particularly accurate. They concluded that studies that are considered high quality (i.e., valid and reliable) provide better outcomes of agreeability between collegiate student and content expert (Falchikov & Boud, 1989).

Several studies in the field of music teacher education, specifically, support the findings that, through the use of definitive self-evaluation tools that provide criteria and operational definitions, students assess their teaching performances accurately, and therefore, self-assessment can be used as a reliable form of feedback on performance achievement (Alley, 1980; Yarbrough, 1987). Specifically, studies found that by videotaping teaching episodes, students could execute an accurate self-assessment when using a straightforward and evaluative tool to analyze their performance and, "…it was possible to obtain a quantitative profile of a teacher's use of various behaviors…[which] could then be pinpointed and, conceivably, modified" (Rosenthal, 1985, p. 18). Yarbrough et al. (1979) used assessment forms constructed in a previous study in order to compare students' self-assessment feedback, versus the traditional mode of feedback from a content expert, when observing video recorded conducting episodes. They found that students in the self-assessment group did just as well as students who were given

feedback from a content expert, thus there is evidence supporting self-assessment effectiveness from video recorded observations if students utilize a valid assessment tool which is appropriate for the context of their video segment.

The authenticity of the setting, a combination of a range of factors that may influence teacher preparation, is a determining factor for the effectiveness and quality of the potential learning experience in which pre-service educators complete their teacher training (Houston, 2008). Education research, and teacher education research specifically, have been repeatedly criticized for studies of low quality which are not characterized by rigorous methodologies and procedures, are not generalizable, are challenging to replicate due to lack of a disclosed method, and are consequently difficult to draw decisive conclusions from (Floden, 2008). As the level of accountability for student success increases in the United States, the field of music education cannot ignore the continuous push towards standards- and research-based methods of assessment for student growth and "…the field of music is at a clear pivot point where the "subjectivity" of music making is becoming progressively interconnected with the "objectivity" of measuring student achievement and program accountability" (Wesolowski, 2019, p. 502). Therefore, future music educators must be actively and methodically trained during their programs to use substantive measures of formative assessment to inform teaching practices and promote student learning if they are expected to be competent teachers in the 21st century classroom (Asmus, 2000; Deluca & Klingerb, 2010).

Pre-service Teachers' Quality of Formative Verbal Feedback

Music teacher evaluations are often lacking in relevance, rigor, objectivity, and validity (Conway & Hibbard, 2020). Specifically, there is a lack of empirical means as a tool for self-assessment to evaluate teachers' verbal feedback in the instrumental secondary ensemble setting

(Flanders, 1965; Sadler, 1989). *High-quality* feedback is at the center of all student learning (Carless et al., 2011; Crisp, 2007; Hattie & Timperley, 2007; Hounsell, 2003; Mulliner & Tucker, 2017; Nicol & Macfarlane-Dick, 2006; Sadler, 1989, 1998). There is an emphasis on the word *high-quality* because it important to discern the difference between the effectiveness of instructive feedback versus general, surface-level observations that do not produce any meaningful data (Freiberg, 1987). Verbal feedback functions as the primary avenue for formative assessment in the performance-based classroom and must be effective, efficient, appropriate for the context, timely, and individualized (Black & Wiliam, 1998; Sadler, 2010; Shute, 2008). Formative assessment can be defined as an, "assessment undertaken during the process of learning to inform the learner as [he or] she moves from some current level of capacity toward mastery of an intended learning outcome" (Brookhart, 2017, p. 927). Sadler (1998) takes the definition a step further saying formative assessment, "…refers to assessment that is specifically intended to provide feedback on performance to improve and accelerate learning" (p. 77). When teachers use formative assessment to examine student performance in the ensemble setting, they must make what Sadler (2010) calls "qualitative human judgements" in order to deliver the most effective verbal feedback to the student. These judgements are particularly difficult in the performance-based classroom because music performance is not simply correct or incorrect. Feedback practices have evolved in modern educational environments and have moved away from the teacher-centered mode of feedback, where information is solely transmitted from teacher to student. High-quality feedback is now viewed as a student-centered practice that supports the processes of students' current and future learning goals (Mulliner & Tucker, 2017).

Purpose

The purpose of this study was to examine the self-assessment accuracy of pre-service music educators' quality of verbal feedback in the context of secondary-level instrumental ensemble instruction using the *Pre-service Music Teacher Verbal Feedback Evaluation Scale*. The research questions that guided this study include:

1. Overall, how accurate were pre-service music educators' perceptions of their verbal feedback when compared to content experts' perceptions?
2. How does accuracy vary across each item of the scale?
3. How does accuracy vary across each domain of the scale?

Method

Measurement Instrument

The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* (see Figure 2.1) was the measurement instrument used in this study. The rating scale was validated, and psychometric considerations were examined, in a previous study using the Many Facet Rasch Partial Credit (MFR-PC) model and results demonstrated that the scale criteria (e.g., scale items) had a high reliability measure (*Rel* = .97). Criteria and domains were carefully constructed based upon the research literature on verbal feedback in performance-based settings, and were screened for clarity, coherence, and readability by the authors and three content experts in the field of music teacher education. The scale structure is based on a four-point Likert-type scale (e.g. *disagree* to *agree*) and includes 35 criterion across five domains: (1) context/audience (n = 10); (2) learning objective/focus (n = 8); (3) timing (n = 4); (4) type (n = 9); and (5) tone (n = 4). Response anchors were dependent upon the objective of each individual criteria, resulting in three sets of anchors: (a) frequency (*never, rarely, sometimes, often*), (b) agreeability (*strongly disagree,*

disagree, agree, strongly agree), and (c) appropriateness (*inappropriate, slightly inappropriate, slightly appropriate, appropriate*). The original scale included 39 items and was revised after a post hoc examination of each item and domain using infit statistics resulting from the Many Facet Rasch Partial Credit (MFR-PC) model. Four items were found to have inconsistent response patterns and demonstrated poor infit, indicating the items were not reliable and were therefore removed from the scale before its use in this study. When both pre-service and in-service teachers are evaluated by professors or administrators, they typically only observe a snapshot of a lesson and it may only occur a handful of times during the school year (Freiberg, 1987). Because of this, it was critical that each criterion could be answered from observing only a short teaching segment, averaging about 10 minutes. The remaining 35 criteria of the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* were observable during these short teaching segments.

Participants

Ratings for this accuracy analysis using the Multifaceted Rasch Rater Accuracy (MFR-RA) measurement model consisted of two groups of raters: operational raters ($N = 44$) and expert raters ($N = 5$). The operational rater participants were pre-service music educators from two large universities in the United States. They were enrolled in their final semester of music education coursework and were preparing to complete their student teaching experiences the following semester. This study was voluntary, and the pre-service music teachers (i.e., operational raters) were informed of the details of the study and required to give written consent. Each pre-service teacher submitted a video recorded teaching segment lasting about 10 minutes in length ($M =$ 11.44 minutes, $SD = 0.96$) of their teaching while participating in authentic field experiences such as practicum teaching at surrounding schools or lab teaching in their instrumental methods

courses. They were asked to watch their teaching segment and assess their ability to give high-quality verbal feedback during a secondary instrumental ensemble rehearsal using the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale*. It was important for the pre-service teachers to re-watch their teaching segment prior to completing the rating scale, so they were not trying to recall details of their performance from memory.

The content experts (i.e., expert raters) were pre-service music teacher educators in the field of secondary instrumental music education and had experience in mentoring pre-service teachers during their training programs. Therefore, due to the extensive experience of the five content experts, the accuracy of pre-service teachers' self-assessment responses when evaluating the quality of their verbal feedback was compared against the evaluation of these expert raters. Each music content expert was assigned a group of pre-service teachers and asked to evaluate their verbal feedback during video recorded teaching segments. Both the content expert and pre-service teacher used the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* for their evaluation. Neither the pre-service teachers nor content experts were explicitly trained in quality verbal feedback measures before completing the rating scale assessment; however, they had access to the scale and were asked to review and become familiar with the criteria before evaluating the teaching segment. Pre-service teachers (i.e., operational raters) and teacher educator content experts (i.e., expert raters) watched the entirety of a teaching segment before completing the evaluation. In order to show high accuracy achievement for the criteria on the scale, the pre-service teacher must have responded to any criteria in an identical manner to their assigned content expert.

The Requirements of Invariant Measurement in Rater-mediated Assessments

The Rasch measurement model has requirements of which are called invariant measurement. Specifically, the requirements of invariance for rater-mediated assessments are as follows: (a) The measurement of the pre-service teacher (i.e., operational rater) must be independent of the particular content expert (i.e., expert rater) that happens to be used for the measuring: rater-invariant measurement of persons; (b) A more able pre-service teacher must always have a better chance of achieving higher accuracy when compared to content experts than a less able pre-service teacher: non-crossing person response functions; (c) The calibration of the content experts must be independent of the particular pre-service teachers used for the calibration: person-invariant calibration of raters; (d) Any pre-service teacher must have a better chance of obtaining a higher rating from lenient content experts than from more severe content experts: non-crossing rater response functions; and (e) Pre-service teachers and content experts must be simultaneously located on a single underlying latent variable: variable map (see Figure 2) (adapted from Engelhard, 2013; Engelhard & Wind, 2018). The use of the Multifaceted Rasch Rater Accuracy (MFR-RA) measurement model allows for the examination of the accuracy of pre-service teachers' self-assessment ratings when compared to the ratings of a content expert. The overall accuracy of the pre-service teachers' self-assessments intimated the expectations of the MFR-RA measurement model. This suggests that the pre-service teachers' self-assessments, the scale criterion, and the scale domains were proportionately accurate for the particular sample of pre-service music teachers examined in this study.

The Multifaceted Rasch Rater Accuracy Measurement Model

Rasch measurement belongs to the family of Item Response Theory (IRT) measurement models and allows researchers to create tools to measure psychological constructs by examining

secondary behaviors (i.e., inferred behaviors). In Rasch measurement, these constructs, also called latent variables, are not directly observable and therefore require the measurement of secondary behaviors using a carefully constructed set of descriptive criteria of the latent variables being examined (Wesolowski, 2019). Rasch measurement theory has requirements of invariance which, when adequate data-to-model fit is achieved, allows latent variables to be mapped onto a continuum which is unidimensional (Engelhard, 2013; Linacre, 2002; see also Figure 2.2 Variable map)

The Multifaceted Rasch Rater Accuracy (MFR-RA) measurement model was adapted from the Many Facet Rasch (MFR) measurement model (Engelhard, 1996; Linacre, 1989/1994). The accuracy model utilized in the data analysis for this study examined ratings from two types of raters: (a) operational raters, and (b) expert raters (Engelhard, 1996). Expert raters are considered content experts in their field and their responses were used as model ratings. Operational raters' responses were compared to the responses of the expert raters and examined for accuracy (i.e., the rating of the operational rater matched that of the expert rater). Engelhard (1996) provides the following mathematical formula for a dichotomous (i.e., 1 = correct response; 0 = incorrect response) MFR-RA measurement model:

$$\ln\left[\frac{P_{ij(x=1)}}{P_{ij(x=0)}}\right] = \lambda_i - \beta_j, \tag{1}$$

where

$P_{ij(x=1)}/P_{ij(x=0)}$ = the probability that rater *i* provides an accurate rating $(\chi = 1)$, rather than an inaccurate rating $(\chi = 0)$ when assessing the teaching segment *j*;

λ_i = the ability of rater *i* to provide accurate ratings; and

β_j = the difficulty associated with providing an accurate rating when assessing the teaching segment *j*.

Three additional facets were added to Equation 1 which represent rater accuracy across raters, rater accuracy across criteria, and rater accuracy across domains. The new equation represents the three facets evaluated in the MFR-RA model (Engelhard, 1996). Equation 2 was implemented using the *FACETS* computer software program (Linacre, 2014):

$$\text{In}\left[\frac{P_{ijmk(x=1)}}{P_{ijmk(x=0)}}\right] = \lambda_i - \beta_j - \delta_m - \eta_k, \tag{2}$$

where

$P_{ijmk(x=1)}/P_{ijmk(x=0)}$ = the probability that rater *i* provides an accurate rating $(\chi = 1)$, rather than an inaccurate rating $(\chi = 0)$ when assessing the teaching segment *j* on criteria *m* within domain *k*;

λ_i = the ability of rater *i* to provide accurate ratings; and

β_j = the difficulty associated with providing an accurate rating when assessing the teaching segment *j*;

δ_m = the difficulty associated with providing an accurate rating on criteria *m*;

η_k = the difficulty associated with providing an accurate rating on domain *k*.

Equation 2 was utilized to map the location of pre-service teachers, criteria, and domains on a linear scale using their logit-score. The linear scale demonstrates the variation of assessment accuracy for the three facets (Engelhard, 1996; Wind & Engelhard, 2013; Wolfe at al., 2016).

Results

Summary Statistics

Table 2.1 provides the summary statistics for the three facets examined in the MFR-RA measurement model. A chi-squared test of significance (χ^2) and reliability of separation (*Rel*) statistics were used as measures to evaluate if there was a significant statistical difference between the pre-service teachers' responses and the content experts' responses. Results suggest that overall, pre-service teachers perceive their ability to give effective verbal feedback during ensemble teaching accurately when compared to their content experts' perceptions. The measurement report for the pre-service teachers' recorded teaching segments demonstrated a low measure of separability, ($\chi^2_{(44)} = 71.2$, $p < .001$, $Rel = .51$), suggesting little difference (i.e., high accuracy) in the way pre-service teachers evaluated themselves compared to content experts. The item measurement report for the scale demonstrated a low measure of separability, ($\chi^2_{(35)} = 72.0$, $p < .001$, $Rel = .66$), suggesting marked differences in accuracy across criterion. The domain measurement report demonstrated a high separability measure, ($\chi^2_{(5)} = 18.1$, $p < .001$, $Rel = .84$), suggesting marked differences in accuracy across domains. All students, criteria, and domains demonstrated a characteristic fit to the MFR-RA measurement model using the parameter of 0.60-1.40 as an indicator of model-to-data fit and results suggest strong construct and predictive validity (Linacre, 2002).

Variable Map

Figure 2.2 shows the variable map for the MFR-RA measurement model, which is a visual representation of the unidimensionality of the latent variables being examined, which satisfy the requirements of invariant measurement (Engelhard, 2013). The variable map represents the operational definition of the latent construct*: accuracy of pre-service music*

educators' self-assessment of the quality of their verbal feedback. Specifically, the results of the analysis are displayed in columns, where each column demonstrates the spread in variability of the elements for each facet included in the model. The first column provides the log odds measure for each facet. Column 2 provides the spread of accuracy achievement for the students (i.e., pre-service music educators), ranging from low self-assessment accuracy (e.g., lower log odds, bottom of the column) to high self-assessment accuracy (e.g., higher log odds, top of the column). Column 3 provides the spread of criteria difficulty, ranging from less difficult criteria to accurately assess (e.g., lower log odds, bottom of the column) to more difficult criteria to accurately assess (e.g., higher log odds, top of the column). Column 4 provides the spread of domain difficulty, ranging from least difficult domain to achieve assessment accuracy (e.g., lower log odds, bottom of the column) to most difficult domain to achieve assessment accuracy (e.g., higher log odds, top of the column).

Pre-service Raters' Overall Self-Assessment Accuracy

Research question 1 (*Overall, how accurate were pre-service music educators' perceptions of their verbal feedback when compared to content experts' perceptions?*) examined how accurately pre-service music educators (i.e., operational raters) evaluated their video recorded teaching segment in comparison to how a content expert (i.e., expert rater) evaluated the same segment. Pre-service teachers who displayed a higher mean logit rating demonstrated a greater ability to achieve self-assessment accuracy, whereas teachers displaying a lower mean logit score demonstrated a lesser likelihood of achievement in self-assessment accuracy. Table 2 provides the calibration of the student facet (i.e., pre-service teachers) and logit-score locations for the 44 pre-service music teacher participants' self-assessment accuracy scores. Scores ranged from the most accurate teacher rater, teacher 23 (average accuracy score = 0.57) with an overall

rating of 0.42 logits (*SE* = 0.35), to the least accurate teacher rater, teacher 31 (average accuracy score = 0.14) with an overall rating of -1.88 logits (*SE* = 0.51). Results show the pre-service teacher raters have good model-to-data fit for the MFR-RA measurement model. A visual representation of Table 2.2 can be found in column 2 of the variable map (see Figure 2).

Pre-service Raters' Accuracy Across Criterion

Research question 2 (*How does accuracy vary across each criterion of the scale?*) addressed the question of which criteria were more or less likely for pre-service music teachers to achieve self-assessment accuracy in comparison to their content expert. Overall, criteria with a higher mean score were more difficult for pre-service teachers to assess accurately, while a lower mean score indicated the criteria was easier to assess accurately. When reviewing the variable map, it is important to note the difference in interpretations of column 2 and column 3. A high score in column 2 denotes higher self-assessment accuracy, whereas a higher score in column 3 denotes criteria that are more difficult to accurately achieve. A low score in column 2 denotes lower self-assessment accuracy, whereas a lower score in column 3 denotes criteria that are easier to accurately achieve. Criteria 2.02 (*The feedback administered focuses on the goals (e.g., teaching focus) of the lesson.*) had a rating of 1.06 logits (*SE* = 0.42) and was the item least likely to be rated accurately by pre-service teachers compared to the assessment of a content expert (average accuracy score = 0.16). Criteria 5.00 (*Teacher communicated with students in a respectful tone.*) had a rating of -2.23 logits (*SE* = 0.53) and was the item most likely to be rated accurately by pre-service teachers compared to the assessment of a content expert (average accuracy score = 0.91). Table 2.3 provides the calibration of criteria facet and logit-score locations for the 35 scale criteria. A visual representation of Table 3 can be found in column 3 of the variable map (see Figure 2.2).

Pre-service Raters' Accuracy Across Domains

Research question 3 (*How does accuracy vary across each domain of the scale?*) addressed the question of how likely it was for pre-service music teachers to achieve self-assessment accuracy across each domain of the scale in comparison to their content expert. Table 4 provides the calibration of domains facet and logit-score locations for the five domains of the scale. Domain 2 (*Learning Objective/Focus*) had a rating of 0.32 logits (*SE* = 0.12) and was the domain least likely to be rated accurately by pre-service teachers compared to the assessment of a content expert (average accuracy score = 0.31). Domain 5 (*Tone*) had a rating of -0.57 logits (*SE* = 0.18) and was the domain most likely to be rated accurately by pre-service teachers compared to the assessment of a content expert (average accuracy score = 0.63). A visual representation of Table 2.4 can be found in column 4 of the variable map (see Figure 2.2).

Conclusion

The purpose of this study was to examine the self-assessment accuracy of pre-service music educators' quality of verbal feedback in the context of secondary-level instrumental ensemble instruction using the *Pre-service Music Teacher Verbal Feedback Evaluation Scale*. The first research question sought to examine the overall accuracy of the pre-service teachers' self-assessment. The results reported a range of scores from 0.42 logits (*SE* = 0.35) for pre-service teacher 23 (average accuracy score = 0.57) to -1.88 logits (*SE* = 0.51) for pre-service teacher 31 (average accuracy score = 0.14). The second research question sought to examine the accuracy of the pre-service teachers' perceptions across each criterion of the scale. The results reported a range of scores from 1.06 logits (*SE* = 0.42) for Criteria 2.02 (average accuracy score = 0.16) to 1.06 logits (*SE* = 0.42) for Criteria 5.00 (average accuracy score = 0.91). The third research question sought to examine the accuracy of the pre-service teachers' self- assessment

across each domain of the scale. The results reported a range of scores from -0.57 logits (*SE* = 0.18) for Domain 2 (average accuracy score = 0.31) to -0.57 logits (*SE* = 0.18) for Domain 5 (average accuracy score = 0.63).

Discussion

As a profession, we must be cognizant of the importance of empirical data-driven assessments. Natriello (1987) performed a thorough investigation on the impact that evaluation has on students in elementary and secondary school classrooms and examined the apparent confusion surrounding the purpose of evaluation in the classroom, as he finds they are often non-descriptive, do not take in to account the influence of multiple unforeseen factors, and fail to consider the multiple purposes for which evaluation should be utilized in the classroom. He recommends further research be performed to examine what methods are being used in schools, but stresses that assessment at the fundamental level requires specified tasks, criteria to be examined based on set standards, performance exemplars, appraisal, and feedback if teachers are using assessment outcomes for student learning interventions.

The accuracy of these pre-service teachers' perceptions of their feedback quality suggests the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* could be used as a helpful tool to facilitate self-assessment and growth of pre-service music educators during in-class lab and field experience teaching. Although this is not an assessment of students in the secondary classroom, the information resulting from accurate self-assessment of instructor verbal feedback directly affects students. The purpose of teacher self-assessment is to inform teaching practices and valid and reliable tools such as the one utilized in this study are a vital component of good teaching and good learning outcomes.

The authors hope to embed these types of meaningful and valid self-assessments in the pre-service music teacher curriculum which, in lay terms, can be used to "teach teachers how to teach" through the use of self-assessment to inform, identify and modify teaching practices that promote quality teaching and therefore quality learning (Brookhart, 2017; Rosenthal, 1985) Authors would like to use the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* as a benchmark in teacher training to examine pre-service music educators' progress in their administration of high-quality verbal feedback at different points in their degree program. Pre-service teachers cannot be completely prepared when they leave their training programs, as this is an impossible expectation, but teacher educators can provide their students a "toolbox" of resources of which they can call upon for self-improvement strategies, to help further their pedagogical knowledge, and as a result they will be exceptional educators, who are self-regulated learners that take accountability for their own professional growth.

Table 2.1

Summary Statistics from the Multifaceted Rasch Rater Accuracy Model

	Pre-Service Teachers (λ)	**Items (δ)**
Logit-Scale Location		
M	-0.38	0.00
SD	0.53	0.57
N	44	35
Infit MSE		
M	0.99	1.00
SD	0.10	0.12
Std. Infit		
M	0.10	-0.10
SD	0.70	1.20
Outfit MSE		
M	1.00	1.00
SD	0.18	0.17
Std. Outfit		
M	0.10	-0.10
SD	0.90	1.20
Separation Statistics		
Reliability of Separation	0.51	0.66
Chi-Square	71.2*	72.0*
Degrees of Freedom	43	34

Note. *$p < .01$.

Table 2.2

Calibration of Student Facet

Teaching Segment	Average Accuracy Score	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
23	0.57	0.42	0.35	1.00	0.00	0.99	0.00
28	0.57	0.42	0.35	1.01	0.10	1.00	0.00
7	0.54	0.29	0.35	0.89	-1.30	0.85	-0.90
17	0.54	0.29	0.35	1.02	0.30	1.02	0.10
39	0.54	0.29	0.35	1.02	0.30	1.02	0.10
18	0.51	0.17	0.35	1.02	0.30	1.01	0.10
29	0.51	0.17	0.35	0.96	-0.50	0.93	-0.40
24	0.49	0.04	0.35	1.03	0.30	1.03	0.20
1	0.46	-0.08	0.35	0.91	-0.90	0.87	-0.90
2	0.46	-0.08	0.35	0.94	-0.50	0.91	-0.60
8	0.46	-0.08	0.35	1.22	2.30	1.53	3.20
10	0.46	-0.08	0.35	0.95	-0.50	0.94	-0.40
11	0.46	-0.08	0.35	0.97	-0.30	0.97	-0.10
15	0.46	-0.08	0.35	1.11	1.20	1.10	0.60
22	0.46	-0.08	0.35	0.97	-0.20	0.94	-0.30
25	0.46	-0.08	0.35	1.04	0.50	1.01	0.10
30	0.46	-0.08	0.35	1.22	2.30	1.53	3.20
32	0.46	-0.08	0.35	0.95	-0.50	0.94	-0.40
33	0.46	-0.08	0.35	0.97	-0.30	0.97	-0.10
37	0.46	-0.08	0.35	1.11	1.20	1.10	0.60
14	0.43	-0.21	0.36	1.05	0.50	1.29	1.80
36	0.43	-0.21	0.36	1.05	0.50	1.29	1.80
41	0.43	-0.21	0.36	0.98	-0.10	0.97	-0.10
21	0.40	-0.34	0.36	1.00	0.00	0.97	-0.10
44	0.40	-0.34	0.36	0.97	-0.20	0.93	-0.40
5	0.37	-0.47	0.37	1.00	0.80	1.21	1.20
6	0.37	-0.47	0.37	0.92	-0.50	0.89	-0.60
27	0.37	-0.47	0.37	1.02	0.20	1.01	0.10
40	0.37	-0.47	0.37	0.99	0.00	1.05	0.30
3	0.34	-0.61	0.37	0.97	-0.10	0.94	-0.20
13	0.34	-0.61	0.37	0.98	0.00	0.94	-0.20
26	0.34	-0.61	0.37	0.97	-0.10	0.92	-0.40
35	0.34	-0.61	0.37	0.98	0.00	0.94	-0.20
43	0.34	-0.61	0.37	1.13	0.90	1.17	0.90
20	0.31	-0.75	0.38	1.04	0.20	1.02	0.10
42	0.31	-0.75	0.38	1.10	0.60	1.11	0.50
4	0.29	-0.90	0.39	0.97	0.00	0.96	-0.10
12	0.29	-0.90	0.39	0.93	-0.30	0.91	-0.30

19	0.29	-0.90	0.39	1.01	0.10	1.05	0.30
34	0.29	-0.90	0.39	0.93	-0.30	0.91	-0.30
16	0.23	-1.24	0.42	0.88	-0.40	1.06	0.30
38	0.23	-1.24	0.42	0.88	-0.40	1.06	0.30
9	0.14	-1.88	0.51	0.73	-0.70	0.49	-1.30
31	0.14	-1.88	0.51	0.73	-0.70	0.49	-1.30
Mean	0.40	-0.38	0.37	0.99	0.10	1.00	0.10
SD	0.10	0.53	0.40	0.10	0.70	0.18	0.90

Note. Students are arranged from high to low (e.g., highest assessment accuracy to lowest assessment accuracy).

Table 2.3

Calibration of Criteria

	Average Accuracy Score	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
Criteria							
2.02	0.16	1.06	0.42	1.10	0.40	1.36	1.00
1.02	0.20	0.91	0.38	1.04	0.20	1.02	0.10
5.04	0.34	0.90	0.33	0.87	-1.10	0.82	-1.20
2.08	0.23	0.60	0.37	0.98	0.00	0.90	-0.30
4.03	0.27	0.60	0.35	1.24	1.50	1.44	1.90
2.01	0.25	0.47	0.36	0.99	0.00	0.95	-0.10
1.06	0.32	0.29	0.33	1.02	0.10	1.00	0.00
1.08	0.32	0.29	0.33	0.92	-0.60	0.88	-0.70
4.06	0.34	0.26	0.33	0.98	-0.10	0.95	-0.20
4.09	0.34	0.26	0.33	1.06	0.50	1.02	0.10
3.01	0.36	0.19	0.32	1.04	0.40	1.01	0.10
3.05	0.36	0.19	0.32	0.97	-0.20	0.95	-0.30
1.01	0.34	0.18	0.33	1.01	0.10	0.98	0.00
1.10	0.34	0.18	0.33	0.92	-0.60	0.88	-0.80
4.04	0.36	0.16	0.32	0.92	-0.80	0.88	-0.90
2.07	0.32	0.12	0.33	1.01	0.10	1.00	0.00
3.03	0.39	0.09	0.32	0.89	-1.20	0.84	-1.30
1.03	0.36	0.08	0.32	0.87	-1.30	0.83	-1.30
2.03	0.34	0.01	0.33	0.93	-0.60	0.89	-0.70
2.04	0.34	0.01	0.33	0.97	-0.20	0.94	-0.30
1.07	0.39	-0.03	0.32	0.94	-0.60	0.90	-0.80
4.08	0.41	-0.04	0.32	1.26	2.90	1.28	2.40
4.10	0.41	-0.04	0.32	0.92	-1.00	0.88	-1.00
1.05	0.41	-0.13	0.32	0.90	-1.20	0.87	-1.10
2.06	0.39	-0.20	0.32	0.85	-1.70	0.80	-1.70
5.01	0.61	-0.29	0.32	1.18	1.60	1.19	1.60
4.02	0.48	-0.34	0.31	1.06	0.80	1.04	0.50
5.03	0.64	-0.39	0.32	1.29	2.40	1.31	2.30
2.02	0.43	-0.39	0.31	0.93	-0.80	0.91	-0.80
1.04	0.48	-0.42	0.31	0.85	-2.20	0.83	-2.00
4.01	0.50	-0.43	0.31	0.88	-1.70	0.87	-1.60
4.07	0.52	-0.53	0.31	1.05	0.70	1.09	1.10
3.02	0.55	-0.60	0.31	0.98	-0.20	0.98	-0.20
1.09	0.57	-0.81	0.31	1.23	2.50	1.27	2.60
5.00	0.91	-2.23	0.53	1.13	0.40	1.40	0.80
Mean	0.40	0.00	0.33	1.00	-0.10	1.00	-0.10

SD	0.14	0.57	0.04	0.12	1.20	0.17	1.20

Note. The criteria are arranged from high to low (e.g., most accurate to least accurate).

Table 2.4

Calibration of Domains

	Average Accuracy Score	**Measure**	***SE***	**Infit *MSE***	**Std. Infit *MSE***	**Outfit *MSE***	**Std. Outfit *MSE***
Domain							
Learning Objective/Focus	0.31	0.32	0.12	0.96	-0.80	0.97	-0.40
Context/Audience	0.37	0.15	0.10	0.97	-0.90	0.94	-1.10
Type	0.40	0.07	0.11	1.04	1.20	1.05	1.10
Timing	0.41	0.04	0.16	0.97	-0.60	0.95	-0.90
Tone	0.63	-0.57	0.18	1.12	1.50	1.18	1.20
Mean	0.42	0.00	0.13	1.01	0.10	1.02	0.00
SD	0.11	0.30	0.03	0.06	1.10	0.09	1.20

Note. The domains are arranged from high to low (e.g., most difficult to achieve accuracy to least difficult to achieve accuracy).

Domain	Rating Scale Categories			
Context/Audience				
1.01. Teacher provides individual feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.02. Teacher provides feedback to small groups of students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.03. Teacher provides feedback to the entire ensemble.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.04. Teacher chooses the appropriate method of feedback (e.g., individual or group) for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.05. The chosen type of feedback (e.g., oral, demonstration, nonverbal) is appropriate for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.06. Teacher addresses student errors during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.07. Teacher addresses student misconceptions during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.08. Teacher checks for student understanding of feedback.	Never	Rarely	Sometimes	Often
1.09. Teacher appropriately fields questions/responses from their students about feedback given.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.10. Teacher provides enough feedback for students to understand the objective.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Learning Objective/Focus				
2.01. Teacher either states, or makes clear through their feedback, what the learning objective is for the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.02. The feedback administered focuses on the goal (e.g., teaching focus) of the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.03. Teacher provides feedback about the learning process.	Disagree	Somewhat Disagree	Somewhat Agree	Agree

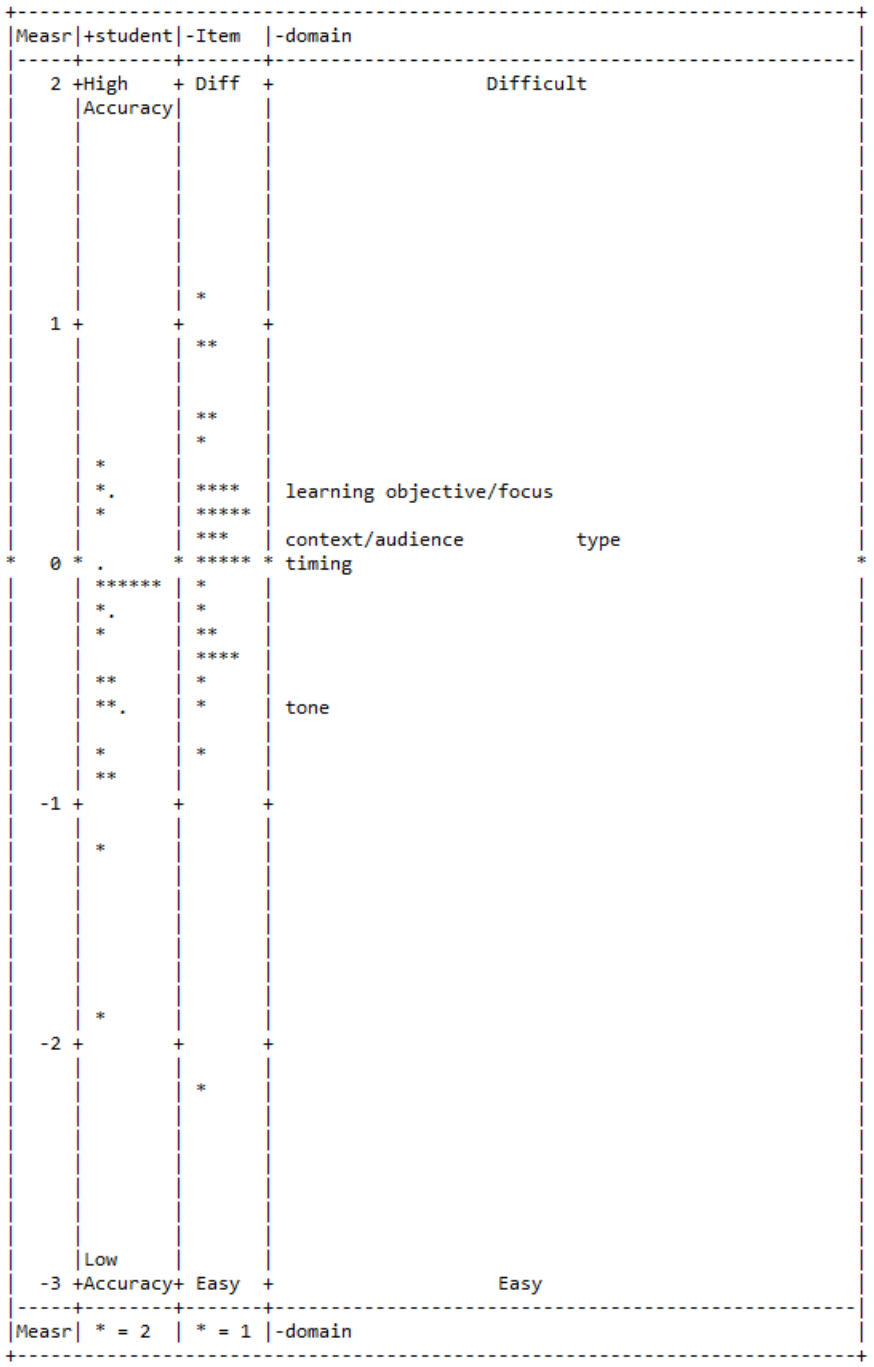
2.04. Teacher compares (by administering feedback) student work to established criteria.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.05. Teacher compares (by administering feedback) student work to past performances.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.06. Teacher provides steps for improvement when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.07. Teacher makes a connection to students' prior knowledge when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.08. Teacher makes a connection to real life situations when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Timing				
3.01. The timing of teacher feedback is appropriate for the event being addressed (e.g. it is given while students are still mindful of the learning target).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.02. Feedback is provided at an appropriate time (e.g., feedback relates to the learning task at hand).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.03. Teacher takes advantage of teachable moments to give feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
3.05. Students are presented opportunities to apply feedback once it is given.	Never	Rarely	Sometimes	Often
Type				
4.01. Teacher provides oral feedback.	Never	Rarely	Sometimes	Often
4.02 Teacher provides feedback by modeling/demonstrating.	Never	Rarely	Sometimes	Often
4.03. Teacher uses nonverbal feedback (e.g., picture, diagrams, gestures).	Never	Rarely	Sometimes	Often
4.04. Teacher chooses appropriate feedback content (e.g., relates to the lesson, or student/ensemble error).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
4.06. Teacher uses feedback to prompt student discussion/reflection.	Never	Rarely	Sometimes	Often

4.07. Teacher compares the performance of a student/group of students to another student/group of students (norm-referenced).	Never	Rarely	Sometimes	Often
4.08. Teacher compares the performance of a student/group of students to a set of standards (criterion-referenced).	Never	Rarely	Sometimes	Often
4.09. Teacher uses self-referenced feedback (e.g., directly compares student performance to their previous performances).	Never	Rarely	Sometimes	Often
4.10. The function of teacher feedback is descriptive in nature.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Tone				
5.00. Teacher communicates with students in a respectful tone (e.g., positive tone quality).	Never	Rarely	Sometimes	Often
5.01. Teacher provides positive comments on student performance.	Never	Rarely	Sometimes	Often
5.03. Teacher uses simple/understandable vocabulary when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
5.04. Teacher is clear when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree

Figure 2.1. Pre-Service Music Teacher Verbal Feedback Evaluation Scale (Revalidated)

Figure 2.2

Variable Map for the Many Facet Rasch Rater Accuracy Model



CHAPTER 4

ATTRIBUTES OF PRE-SERVICE MUSIC EDUCATORS' VERBAL FEEDBACK IN THE SECONDARY-LEVEL INSTRUMENTAL MUSIC CLASSROOM[3]

[3] Athanas, M.I. and B.C. Wesolowski. To be submitted to *Research Studies in Music Education.*

Abstract

The purpose of this study is to examine pre-service teachers' characteristics of their formative verbal feedback. In the first study, a rating scale (*Pre-Service Music Teacher Verbal Feedback Evaluation Scale*) was constructed and validated to examine the quality of pre-service music educators' formative verbal feedback in the secondary-level instrumental music classroom. The second study examined pre-service music educators' self-assessment accuracy of the quality of their formative verbal feedback when compared to the ratings of a content expert. This study will focus on identifying subsets of pre-service students and criteria within the data collected in order to look for commonalities of their teaching behaviors and whether those behaviors are connected in a meaningful way. For example, in study one and two, pre-service music educators found domain two (*learning objective/focus*) to be the most difficult to achieve and to rate their achievement accurately. They found domain 5 (*tone*) the easiest to achieve and rate their achievement accurately. Although these domains were distinguishable as most to least difficult, not all criteria in each of the domains were the most or least difficult. For example, criteria 2.02 (*The feedback administered focuses on the goal (e.g., teaching focus) of the lesson*) in domain two was the most difficult item on the rating scale, but in the same domain, 2.04 (*Teacher compares (by administering feedback) student work to established criteria.*) was a less difficult criteria. A cluster analysis will be performed to identify if there are certain criteria which can be grouped together, as well as if there are pre-service music educator teaching segments with similar attributes that cause their teaching performances to be rated similarly.

Authors hope that information gleaned from this study may positively influence teacher training programs by identifying weaknesses and strengths within the curriculum on pre-service teachers' verbal feedback preparation.

Introduction

The field of education in the United States has recently begun to *systematically* examine the construct of "feedback" in the classroom as a significant tool for teaching, as traditionally many educators believe assigning a grade is an acceptable way to inform students on their performance, failing to separate grading and feedback as two different types of assessment (Hattie, 2019). A study completed by Butler (1988) showed that both high achieving and low achieving students are more interested and intrinsically motivated when they expect to receive formative, task-oriented feedback, such as constructive comments, versus when they expect to simply receive a grade in the form of summative feedback.

Formative assessment using verbal feedback is an essential component of the performance-based learning. Verbal feedback in the performance-based classroom is the primary method used to convey information to students throughout an ensemble rehearsal. Throughout the three studies examining quality pre-service music educator verbal feedback using the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale*, authors consistently used the same definitions to define formative assessment and high-quality verbal feedback. These definitions were derived from D. Royce Sadler, Susan Brookhart, John Hattie, and Valerie Shute, some of the leading influential researchers on feedback and assessment in the field of education. For the purposes of these studies, formative assessment is defined as an "assessment undertaken during the process of learning to inform the learner as [he or] she moves from some current level of capacity toward mastery of an intended learning outcome" (Brookhart, 2017, p. 927). Formative assessment should be used consistently in the performance-based classroom to improve the learning process (Sadler, 1998). Verbal feedback is defined as, "information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of

improving learning" (Shute, 2008, p.154). A crucial, but often overlooked, component of verbal feedback is the examination of the students' understanding and ability to execute the given feedback, and thus, "…there seems [to be] little point in maximizing the amount and nature of feedback given if it is not received and understood" (Hattie, 2019, p. 5). Therefore, in order for feedback to be meaningful, all feedback messages in relation to student performances, strengths and weaknesses, and recommendations for improvement must be presented in relation to a previously established set of exemplars such as learning objectives, standards, and criteria, that exemplify the end-goal of a lesson or unit (Nicol & Macfarlane-Dick, 2006). Finally, the students' ability to process and internalize information is dependent on an important factor that helps define quality classroom instruction: differentiation.

Differentiated Instruction

The term *differentiation* has been one of the primary buzz words in academia for the past two decades, beginning in 1999 when Carol Ann Tomlinson developed a model for teaching using differentiation in the classroom (Tomlinson, 1999). Tomlinson (2014) describes high-quality teachers as, "students of their students… [and] diagnosticians, prescribing the best possible instruction based on both their content knowledge and their emerging understanding of students' progress in mastering critical content… [who] are also artists who use the tools of their craft to address students' needs" (p. 4). Furthermore, high-quality teachers differentiate based upon the students in their classroom. Teachers who have mastered differentiated instruction: (a) believe in their students' abilities to succeed, and understand that each student is an individual who cannot be expected to conform to a specific learning style, (b) promote an inclusive and nurturing classroom environment, (c) are flexible and employ an array of different learning strategies in order to meet their students' needs, (d) promote student success by ensuring their

students are continuously held to high standards as they strive to accomplish designated learning goals, and intrinsic motivation, (e) encourage students to compete with their own past performances, instead of comparing their work to the performances of others, and arguably most importantly, (f) differentiated instruction is accompanied by continuous formative assessment and high-quality feedback and instruction from the teacher, whose job it is to inform students on the strengths and weaknesses of their performance, and to collaborate and offer continual support and guidance throughout the learning cycle (Tomlinson, 2014).

Alton-Lee (2003) provided ten research-based characteristics to be used with a wide variety of student learners of differing ability, knowledge, and age. She clearly defines *quality teaching* as, "pedagogical practices that facilitate for heterogenous groups of students, their access to information, and ability to engage in classroom activities and tasks in ways that facilitate learning related to curriculum goals" (p. 1). The ten criteria derived from an extensive review of the research literature, examining what teacher qualities and characteristics deliver the best student outcomes for a diverse student population (Alton-Lee, 2003; Alton-Lee & Nuthall, 1998). Furthermore, her ten characteristics of quality teaching are as follows:

1. Quality teaching is focused on student achievement (including social outcomes) and facilitates high standards of student outcomes for heterogeneous groups of students.
2. Pedagogical practices enable classes and other learning groupings to work as caring, inclusive, and cohesive learning communities.
3. Effective links are created between school and other cultural contexts in which students are socialized, to facilitate learning.
4. Quality teaching is responsive to student learning processes.
5. Opportunity to learn is effective and sufficient.
6. Multiple task contexts support learning cycles.
7. Curriculum goals, resources including ICT usage, teaching and school practices are effectively aligned.
8. Pedagogy scaffolds and provides appropriate feedback on students' task engagement.
9. Pedagogy promotes learning orientations, student self-regulation, metacognitive strategies and thoughtful student discourse.
10. Teachers and students engage constructively in goal-oriented assessment.

These ten characteristics of quality teaching exemplify all important aspects of the teaching and student learning process and cannot be achieved without a clear understanding of differentiated instruction. Alton-Lee (2003) has provided a systematic tool and invaluable resource for educators to establish high-quality teaching practices, including a breakdown of each of the above characteristics into very descriptive criteria.

Hattie's (2012) principles of feedback to inform learning were influenced by a report defining quality teaching by Alton-Lee (2003) written for the Ministry of Education in New Zealand. Hattie (2019) provides an eight-step process of effective feedback to inform *learning.* Feedback: (a) sparks learning, (b) flourishes in the right environment, (c) clarifies for students where they are going, (d) informs students how they are going, (e) highlights the next steps for improvement, (f) matches the needs of the learner, (g) promotes students' self-regulation, and (h) flows bi-directionally between learners and teachers (p.6).

There are striking similarities between Tomlinson's model for differentiation, Alston-Lee's characteristics of quality teaching, and Hattie's eight-step process to use feedback and formative assessment to inform teaching and learning. These defining characteristics of quality and differentiated instruction are also exemplified in the criteria presented in the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale.* The criteria, established during the construction of a rating scale in a previous study, were largely adapted from common themes found throughout the music education and general education research literature, and have been appropriately adapted to support differentiated, high-quality verbal feedback in the music ensemble classroom. Specifically, Susan Brookhart's (2017) research, rubrics, and recommendations for administering effective, quality feedback in her book *How to Give Effective Feedback to Your Students* was a primary resource influencing the criteria selected for

the pre-service teacher feedback scale. The rating scale specifies detailed criteria in order to examine the quality of pre-service teachers' verbal feedback in the secondary-level instrumental music classroom. Each criterion identifies a different element important to the administration of verbal feedback to students, providing the pre-service music teacher a platform of useful differentiation strategies by context, audience, and students' strengths and weaknesses. Music educators can use differentiated instruction as a primary teaching strategy when they are inevitably presented with groups of students with varying achievement levels and learning styles within the same ensemble.

Purpose

Differential item functioning (DIF) and cluster analyses are tools that have been used in the music education research literature to examine whether meaningful typologies exist within a specific data set (Bernabé-Valero, 2019; Odendaal, 2013, 2016; Wesolowski, 2017; Wesolowski & Wind, 2019; Wesolowski & Ng, 2020; Zhukov, 2007, for example) The purpose of this study was to determine if pre-service music educators could be grouped by common patterns of the quality of their verbal feedback in the music ensemble rehearsal. The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* describes the traits of quality verbal feedback by providing a list of criteria to measure verbal feedback. The research questions that guided this study include:

1. Does a meaningful verbal feedback typology exist based upon systematic differential item functioning (student-by-criterion) bias indices?
2. What are the predominant characteristics of the quality verbal feedback typologies?
3. What can be concluded about the strengths and weaknesses in teacher training programs, if any, from the patterns for typologies of the pre-service teacher?

Method

Participants

The participants examined in this study were a group of secondary-level instrumental pre-service music education students ($N = 55$) at a large American university. Study participation was voluntary and pre-service students were asked to video-record a teaching segment, approximately 10 minutes in length, during an authentic teaching experience. Authentic teaching experiences could include videos from practicum teaching in the surrounding public schools, or lab teaching during instrumental methods courses (e.g., secondary methods, woodwind methods, brass methods, etc.). Teaching video segments were then randomly assigned amongst a group of context expert raters ($N = 15$) to be evaluated using the feedback scale. To be considered a content expert in music education, raters needed to be a practicing in-service teacher, with a minimum of five years teaching experience. Content experts evaluated their assigned teaching episodes using the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale.* Rater responses were then used to perform a differential item functioning analysis (DIF) and a cluster analysis to examine the potential patterns found in the pre-service educators' teaching segments.

Measurement Instrument

This study utilized the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale,* constructed in a previous study, which examined the psychometric qualities (i.e., validity, reliability, and precision) of the scale. The original *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* included criteria ($N = 39$) divided amongst five domains: (a) context/audience ($n = 11$); (b) learning objective/focus ($n = 8$); (c) timing ($n = 5$); (d) type ($n = 11$); and (e) tone ($n = 4$). A four-point Likert-type (see Figure 3.1) rating scale structure was used, and a specific set of Likert scale responses was chosen for each criterion on the rating scale based upon the

substantive interpretation of the criteria. Substantive interpretation of the criteria resulted in three different Likert-type response sets: (a) frequency (*never, rarely, sometimes, often*), (b) agreeability (*strongly disagree, disagree, agree, strongly agree*), and (c) appropriateness (*inappropriate, slightly inappropriate, slightly appropriate, appropriate*). A chi-squared test of significance in the original scale study showed that pre-service teachers could be separated based on their logit-score locations, $\chi^2_{(55)} = 986.30$, $p < .01$, and had a high reliability measure (*Rel* = .94), demonstrating that the spread of pre-service teachers' logit-score locations was significant. A chi-squared test of significance was also conducted to examine if the scale criteria could be separated based on their logit-score locations, $\chi^2_{(39)} = 1000.10$, $p < .01$, and had a high reliability measure (*Rel* = .97), showing that the spread of the criteria logit-score locations was significant.

This scale was constructed with the intention of measuring pre-service music educators' quality of verbal feedback; therefore, the information resulting from the substantive interpretation of the scale scores is only relevant in the context of secondary-level pre-service music educators (Wright & Stone, 1999). Item-mean square (MSE) indices indicate that both the pre-service teaching video segments and the scale criteria displayed good data-to-model fit. A fit indicator of 0.60-1.40 was chosen to identify misfit students and criteria (as this rating scale was not considered a high-stakes assessments) based upon the resulting pattern of responses (Linacre, 2002). The achievement of invariant measurement in the model is also used to interpret good data-to-model fit (Engelhard, 2013). Misfit pre-service teaching segments (students 15, 22, 27, 28, 29, and 44) and misfit scale criteria (1.11, 3.04, 4.05, 4.11, 5.02) were removed from the data set prior to the completion of the differential item functioning (DIF) and cluster analyses.

Psychometric Considerations

Rasch measurement (Rasch, 1960/1980) is under the umbrella of the family of item response theory (IRT). Rasch measurement has five criteria which are required to attain invariant measurement in rater-mediated assessments, and these requirements must be fulfilled in order to achieve the required unidimensionality within the rating scale model (Engelhard, 2013; Engelhard & Wind, 2018). The initial scale study examined the psychometric qualities of the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* using the Many Facet Rasch Partial Credit (MFR-PC) model (Linacre, 1989/1994). The results confirmed that the rating scale had good data-to-model fit, as well as high reliability and validity measures, demonstrating that the rating scale was psychometrically sound in the context of examining pre-service music educators' quality verbal feedback.

In this study, a differential item functioning (DIF) analysis and a cluster analysis were used to examine pre-service music educators' quality of their verbal feedback. Differential item functioning (DIF) can be defined as, "the loss of item estimate invariance across subsamples of respondents" and is also referred to as a bias analysis (Bond & Fox, 2015, p. 359). Therefore, if an item's difficulty varies more than the error within the model, differential item functioning will exist and must be explored for a substantive understanding of the latent variable being examined. Specifically, this study examined the loss of invariance between pre-service music educators' teaching segment scores and the criteria of the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale.* The FACETS (Linacre, 2014) computer software program was used to perform a differential item functioning analysis (i.e., bias analysis) in order to determine if there were meaningful interactions between the pre-service music educators' (i.e., students) verbal

feedback quality and the scale criteria (i.e., items) within the parameters of the Rasch Measurement Model.

R statistics software (R Core Team, 2020) was used to perform a hierarchical cluster analysis, a non-hierarchical k-means cluster analysis, and Squared Euclidian distances, which examined any patterns found in the previously evaluated pre-service video teaching segments and grouped them into "clusters" based upon the similarities of the bias indices deduced from the DIF analysis. A hierarchical cluster analysis was used to explore the options of cluster groups, and specifically employed Ward's Linkage method to examine cluster solutions from 2-10 possible solutions (Ward, 1963). The distance in proximity between the students' criterion scores was evaluated by computing Squared Euclidian distances, which identified and excluded any potential outliers that could affect the proximity between the pre-service teachers being assessed (Romesburg, 1984). Cluster analyses solutions were examined both for their statistical indices and for substantive interpretation. A k-means cluster analysis was performed to ascertain a cluster solution using the pre-specified cluster centroids resulting from the hierarchical cluster analysis. A three-cluster solution resulted from the k-means analysis and was the chosen cluster solution to represent these data. The cluster centroids were used as the anchor means for the cluster analysis and derived from the results of the computed Squared-Euclidian distances.

Results

Research Question 1

Research question 1 examines whether meaningful quality verbal feedback typologies exist based upon systematic differential item functioning bias indices. Specifically, a differential item functioning (DIF) analysis was performed to examine student-by-criterion bias indices to determine possible interaction effects between the two facets. The DIF analysis (i.e., bias

analysis) calculated bias indices for each student participant (i.e., participating pre-service music educators), resulting in 2,145 interaction effects. This calculation was based upon the 55 student participants multiplied by the 39 scale criteria. A data frame was constructed to depict each students' score on each scale criterion. In preparation for analysis, the data frame was rescaled after removing misfit students (ID # 15, 22, 27, 28, 29, and 44) and misfit scale criteria (1.11, 3.04, 4.05, 4.11, 5.02) using min-max scaling [0,1] for each of the three variables in the data frame (student, criteria, and bias size). Cluster analysis examines the distances between the dimensions in the two-dimensional realm after the bias indices are scaled properly using mix-max scaling (Iofee & Szegedy, 2015). A principal components analysis (PCA) was performed to determine the ten scale criteria for Dimension 1 and Dimension 2 that were most heavily weighted within the two-dimensional space, dependent on the values of their Squared Euclidian distances (Figure 3.2). The ten scale criteria identified for Dimension 1 are defined by how the feedback is articulated to the student and how the pre-service teacher responds to the misconceptions or questions resulting from their feedback. The ten scale criteria in Dimension 2 are defined by the primary audience of the feedback message, specifically the differences between individual, small group, and large group feedback. Figure 3.2 provides a visual representation of the ten most heavily weighted criteria contributing to Dimension 1 and Dimension 2. Figure 3.3 provides a visual representation of all scale criteria and where they fall within the two-dimensional space.

Research Question 2

Research question 2 examined possible predominant characteristics of quality verbal feedback to identify typologies of pre-service music educators. A hierarchical k-means cluster analysis was completed, and a three-cluster solution was chosen to represent three typologies of

pre-service teachers' verbal feedback quality. Results showed that there were three types of pre-service music educators separated by predominant characteristics of their verbal feedback, pre-service teachers who were: (a) learning objective and type focused, (b) context/audience, learning process focused, and (c) clarity, goal, and time focused. Teacher A can be defined by their ability to align their lesson to a set of learning objectives, and their ability to differentiate by choosing the appropriate type of feedback based on their students' learning needs. Teacher B can be defined by their ability to determine the proper audience, context, and share the overall process of student learning through quality verbal feedback. Teacher C can be defined by their ability to use clear and understandable terminology and vocabulary, and their ability to administer goal-oriented feedback in a timely manner.

Cluster 1. Cluster 1 encompassed 22.45% ($N = 11$) of pre-service music educators' video teaching segments. The following scale criteria were identified using cluster centroids and can be located in Table 3.1.

1.03. *Teacher provides feedback to the entire ensemble.*
2.01. *Teacher either states, or makes clear through their feedback, what the learning objective us for the lesson.*
2.05. *Teacher compares (by administering feedback) student work to past performances.*
2.08. *Teacher makes a connection to real life situations when giving feedback.*
3.05. *Students are presented opportunities to apply feedback once it is given.*
4.02. *Teacher provides feedback by modeling/demonstrating*
4.03. *Teacher uses nonverbal feedback (e.g., pictures, diagrams, gestures).*
4.06. *Teacher uses feedback to prompt student discussion/reflection.*

Cluster 2. Cluster 2 encompassed 48.98% ($N = 24$) of pre-service music educators' video teaching segments. The following scale criteria were identified using cluster centroids and can be located in Table 3.1.

1.01. *Teacher provides individual feedback to students.*
1.02. *Teacher provides feedback to small groups of students.*
1.04. *Teacher chooses the appropriate method of feedback (e.g., individual or group) for the learning context.*

1.05. *The chosen type of feedback (e.g., oral, demonstration, nonverbal) is appropriate for the learning context.*
1.06. *Teacher addresses student errors during the lesson.*
1.07. *Teacher addresses student misconceptions during the lesson.*
1.08. *Teacher checks for student understanding of feedback.*
1.10. *Teacher provides enough feedback for students to understand the objective.*
2.03. *Teacher provides feedback about the learning process.*
2.07. *Teacher makes a connection to students' prior knowledge when giving feedback.*
3.03. *Teacher takes advantage of teachable moments to give feedback to students.*
4.07. *Teacher compares the performance of a student/group of students to another student/group of students (norm-referenced).*
4.08. *Teacher compares the performance of a student/group of students to a set of standards (criterion-referenced).*
5.01. *Teacher provides positive comments on student performance.*

Cluster 3 Cluster 3 encompassed 28.57% ($N = 14$) of pre-service music educators' video teaching segments. The following scale criteria were identified using cluster centroids and can be located in Table 3.1.

1.09. *Teacher appropriately fields questions/responses from their students about feedback given.*
2.02. *The feedback administered focuses on the goal (e.g., teaching focus) of the lesson.*
2.04. *Teacher compares (by administering feedback) student work to established criteria.*
2.06. *Teacher provides steps for improvement when giving feedback.*
3.01. *The timing of teacher feedback is appropriate for the event being addressed (e.g., it is given while students are still mindful of the learning target).*
3.02. *Feedback is provided at an appropriate time (e.g., feedback relates to the learning task at hand).*
4.01. *Teacher provides oral feedback.*
4.04. *Teacher chooses appropriate feedback content (e.g., relates to the lesson, or student/ensemble error).*
4.09. *Teacher uses self-referenced feedback (e.g., directly compares student performance to their previous performances).*
4.10. *The function of teacher feedback is descriptive in nature.*
5.03. *Teacher uses simple/understandable vocabulary when administering feedback.*
5.04. *Teacher is clear when administering feedback.*

Research Question 3

Research question 3 calls for the examination of the three clusters from a more qualitative standpoint to determine if the cluster typologies represent any patterns of strengths or weaknesses in teacher training programs. The three clusters align with the results of the previous

two studies that utilized the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* to explore pre-service teachers' ability to give quality verbal feedback to their students. In both the validation and revalidation of the verbal feedback rating scale, the domain *learning objective/focus* was the most difficult domain for preservice teachers to achieve during their teaching episodes (see Figure 3.1). We see a similar pattern throughout the three-cluster solution. Cluster 1, the learning objective and type focused group of pre-service teachers, represented 22.45% ($N = 11$) of the pre-service teacher participant sample, separating 77.55% of pre-service teachers due primarily to the amount of difficulty they have aligning their lessons to a learning objective, or their ability to choose the correct type of feedback to administer. Cluster 2, the context/audience, learning process focused group of pre-service teachers, represented the largest participant sample, 48.98% ($N = 24$). Cluster 2 established a group of pre-service music educators who showed strength in their ability to accurately determine the primary audience to direct their verbal feedback on student performance errors and the overall learning process. Cluster 3, the clarity, goal, and time focused group of pre-service teachers, represented 28.57% of the participant sample ($N = 14$). These teachers were clear and timely with their verbal feedback and referred to a set of standards or objectives while teaching. In conclusion, Cluster 2 is represented by the largest number of pre-service teachers, nearly 50% of the entire sample, and symbolizes the criteria which pre-service music educators find easier to achieve. Adversely, Cluster 1 has the smallest number of pre-service teachers represented. This set of criteria were more difficult for pre-service teachers to achieve during their teaching segments. Considering the substantive interpretation of the three clusters, the apparent weaknesses of the evaluated pre-

service teacher participants are telling. Consequences of the strengths and weaknesses of pre-service teachers and teacher training programs will be discussed further in the Discussion section.

Conclusion and Discussion

Objectives-based teaching is crucial to the development of student knowledge in the performance-based classroom, as it allows teachers to construct a specified curriculum which should be clearly communicated, well-defined, appropriate for the student audience, and demonstrates an explicit connection between any particular lesson and the established framework of objectives and expectations (Morrison et al., 2011; Wesolowski, 2015). Quality verbal feedback criteria addressing learning objective-based teaching were consistently some of the most difficult for pre-service music educators to achieve during their teaching episodes, often leaving students unaware of the primary focus of their lesson. Research question 3 focused on understanding the strengths and weaknesses found within the typologies constructed from the patterns of pre-service music educators' ability to administer quality verbal feedback. The results of the substantive interpretation from the cluster analysis are significant and consequential to teacher training programs. Cluster 1 was comprised of the smallest number of pre-service teachers, totaling about 23% of the entire sample. The scale criteria aligned with Cluster 1, from the results of the cluster centroids, list some of the primary skills teacher education programs expect their pre-service teachers to know upon graduation and certification. The skills represented within these criteria show that less than 23% of the pre-service teacher participants in this study were able to consistently tie their verbal feedback to the learning objective they intended to teach during their lesson. The research literature discussed shows that students learn

better when their teachers identify learning objectives, standards, and goals throughout the learning process.

As discussed throughout this study, just because verbal feedback is provided in the right context and to the correct audience does not mean that it is being explicitly aligned to a set of standards, learning objectives, or is meaningful to the student or lesson (Hattie, 2019). The ability for pre-service teachers to choose the correct context and audience for their verbal feedback message seemed to be a definite strength for pre-service teachers; however, that does not mean that said feedback is being directly aligned to a preestablished set of standards, learning objectives, or goals. Furthermore, if pre-service teachers can consistently focus on strengthening their weakest verbal feedback criteria, they can more successfully administer quality verbal feedback, and therefore be more effective educators in the secondary-level instrumental music classroom. Moving forward, music teacher training programs might consider assessing the music education curriculum to evaluate current course offerings, teaching strategies and practices, and the overall curriculum of the teacher training program.

Table 3.1

Finalized Cluster Centroids by Scale Criteria

Cluster	Centroids by Scale Criteria					
	1.01	**1.02**	**1.03**	**1.04**	**1.05**	**1.06**
1	113.36	107.61	105.88	107.90	100.13	107.34
2	90.63	92.54	99.30	89.94	96.42	91.18
3	104.69	108.48	94.27	109.80	105.45	105.70
	1.07	**1.08**	**1.09**	**1.10**	**2.01**	**2.02**
1	102.61	101.36	96.02	100.75	92.81	89.64
2	91.39	98.15	96.85	98.83	101.10	99.30
3	110.99	100.70	109.26	101.22	106.69	114.13
	2.03	**2.04**	**2.05**	**2.06**	**2.07**	**2.08**
1	102.60	91.80	93.16	108.53	99.01	91.20
2	98.33	98.65	102.18	100.01	101.95	103.07
3	100.03	110.52	106.87	91.80	97.72	102.71
	3.01	**3.02**	**3.03**	**3.05**	**4.01**	**4.02**
1	92.22	92.35	97.10	90.19	105.28	93.00
2	98.99	96.24	100.96	104.69	102.26	103.81
3	109.89	116.46	98.31	101.83	91.42	101.40
	4.03	**4.04**	**4.06**	**4.07**	**4.08**	**4.09**
1	90.78	96.73	105.05	94.25	95.31	100.80
2	104.77	96.24	99.30	105.70	106.55	101.56
3	100.13	108.17	96.40	95.75	94.51	95.57
	4.10	**5.01**	**5.03**	**5.04**		
1	100.39	99.89	102.78	109.16		
2	105.20	102.79	103.88	102.44		
3	88.87	94.41	88.89	83.01		

Note. Areas shaded light gray indicate a differentiation from other clusters above 100.00. Areas shaded dark gray indicate differentiation from other clusters below 100.00.

Figure 3.1

Pre-Service Music Teacher Verbal Feedback Evaluation Scale

Domain	Rating Scale Categories			
Context/Audience				
1.01. Teacher provides individual feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.02. Teacher provides feedback to small groups of students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.03. Teacher provides feedback to the entire ensemble.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.04. Teacher chooses the appropriate method of feedback (e.g., individual or group) for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.05. The chosen type of feedback (e.g., oral, demonstration, nonverbal) is appropriate for the learning context.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.06. Teacher addresses student errors during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.07. Teacher addresses student misconceptions during the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
1.08. Teacher checks for student understanding of feedback.	Never	Rarely	Sometimes	Often
1.09. Teacher appropriately fields questions/responses from their students about feedback given.	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
1.10. Teacher provides enough feedback for students to understand the objective.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Learning Objective/Focus				

2.01. Teacher either states, or makes clear through their feedback, what the learning objective is for the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.02. The feedback administered focuses on the goal (e.g., teaching focus) of the lesson.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.03. Teacher provides feedback about the learning process.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.04. Teacher compares (by administering feedback) student work to established criteria.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.05. Teacher compares (by administering feedback) student work to past performances.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.06. Teacher provides steps for improvement when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.07. Teacher makes a connection to students' prior knowledge when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
2.08. Teacher makes a connection to real life situations when giving feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Timing				
3.01. The timing of teacher feedback is appropriate for the event being addressed (e.g., it is given while students are still mindful of the learning target).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.02. Feedback is provided at an appropriate time (e.g., feedback relates to the learning task at hand).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
3.03. Teacher takes advantage of teachable moments to give feedback to students.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
3.05. Students are presented opportunities to apply feedback once it is given.	Never	Rarely	Sometimes	Often
Type				

4.01. Teacher provides oral feedback.	Never	Rarely	Sometimes	Often
4.02 Teacher provides feedback by modeling/demonstrating.	Never	Rarely	Sometimes	Often
4.03. Teacher uses nonverbal feedback (e.g., pictures, diagrams, gestures).	Never	Rarely	Sometimes	Often
4.04. Teacher chooses appropriate feedback content (e.g., relates to the lesson, or student/ensemble error).	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
4.06. Teacher uses feedback to prompt student discussion/reflection.	Never	Rarely	Sometimes	Often
4.07. Teacher compares the performance of a student/group of students to another student/group of students (norm-referenced).	Never	Rarely	Sometimes	Often
4.08. Teacher compares the performance of a student/group of students to a set of standards (criterion-referenced).	Never	Rarely	Sometimes	Often
4.09. Teacher uses self-referenced feedback (e.g., directly compares student performance to their previous performances).	Never	Rarely	Sometimes	Often
4.10. The function of teacher feedback is descriptive in nature.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
Tone				
5.01. Teacher provides positive comments on student performance.	Never	Rarely	Sometimes	Often
5.03. Teacher uses simple/understandable vocabulary when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree
5.04. Teacher is clear when administering feedback.	Disagree	Somewhat Disagree	Somewhat Agree	Agree

Note. Items 1.11, 3.04, 4.05, 4.11, 5.02 removed from scale due to poor infit.

Figure 3.2

Contribution of Scale Criteria to Dimension 1 and Dimension 2 from the Principal Components Analysis

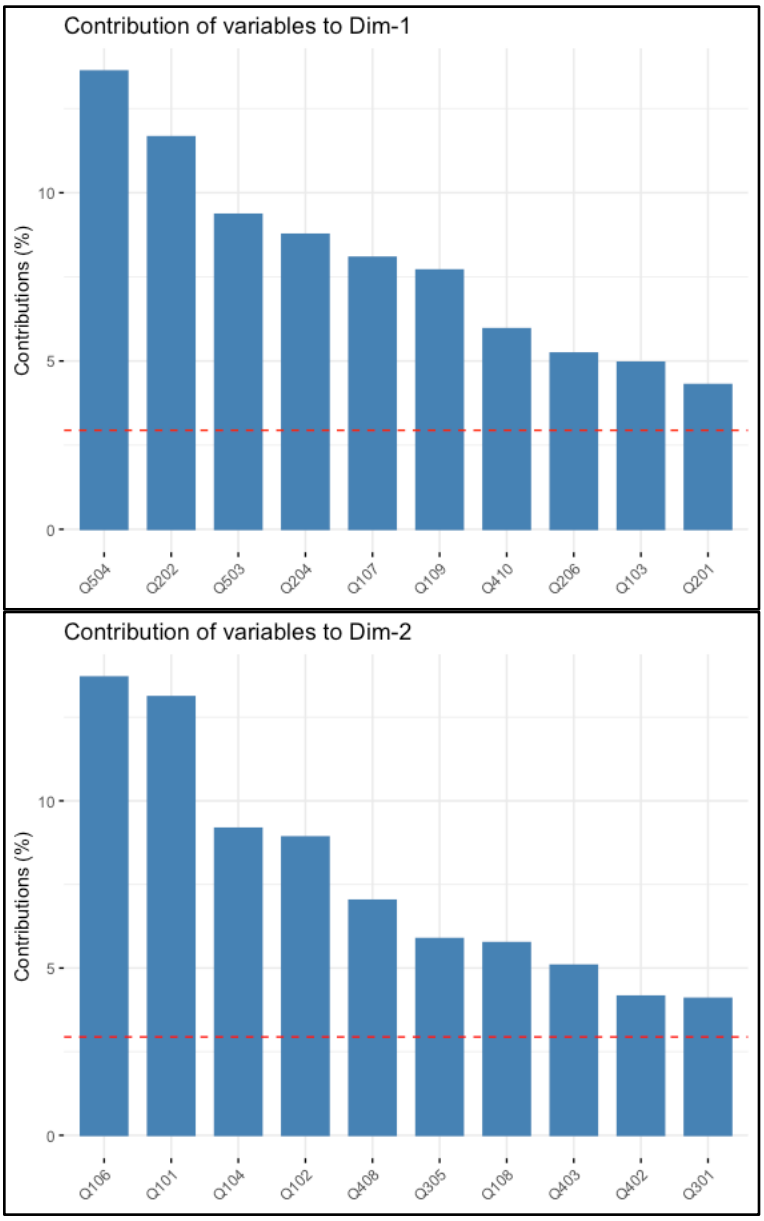


Figure 3.3

Contribution of all Scale Criteria in the Two-Dimensional Space

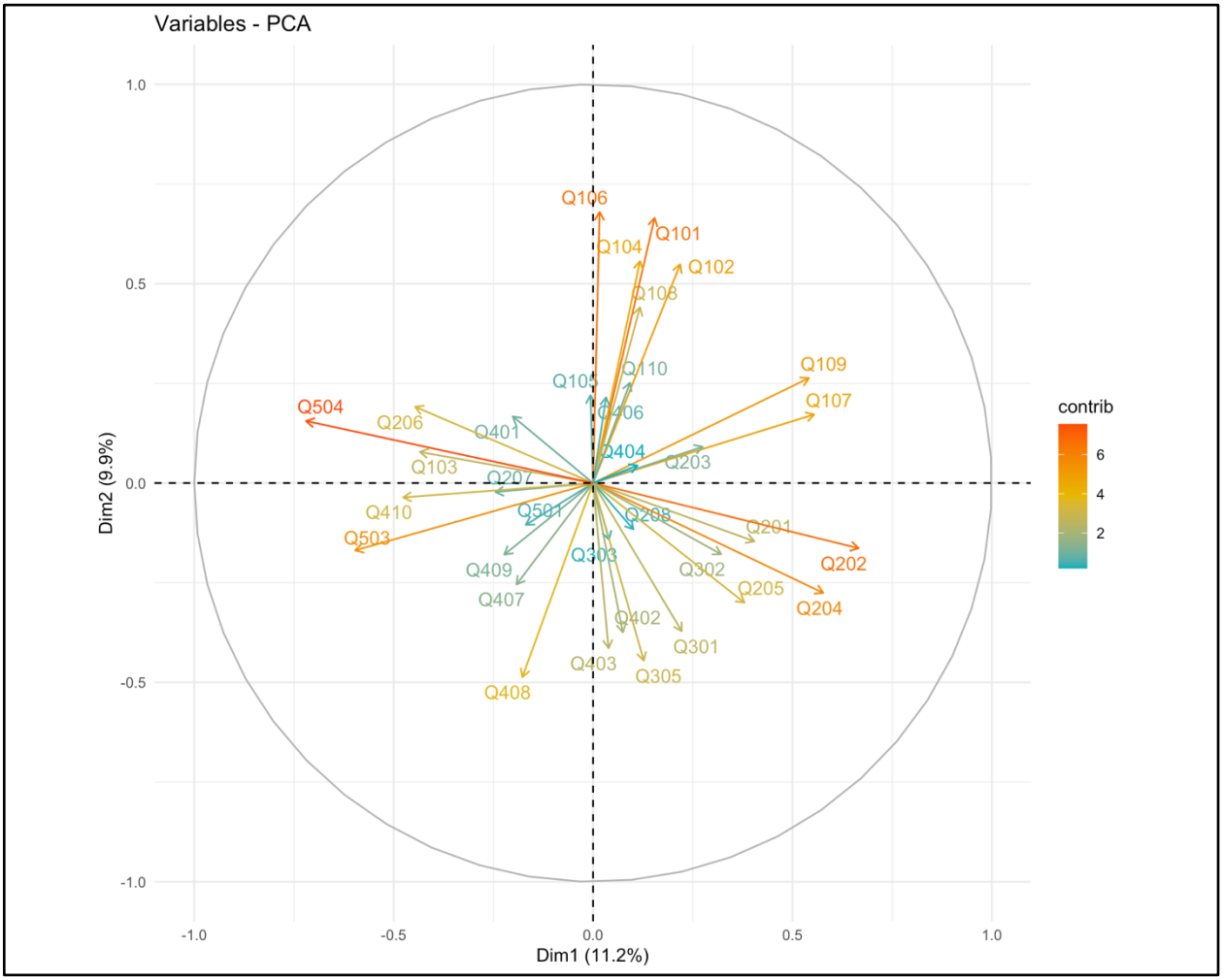
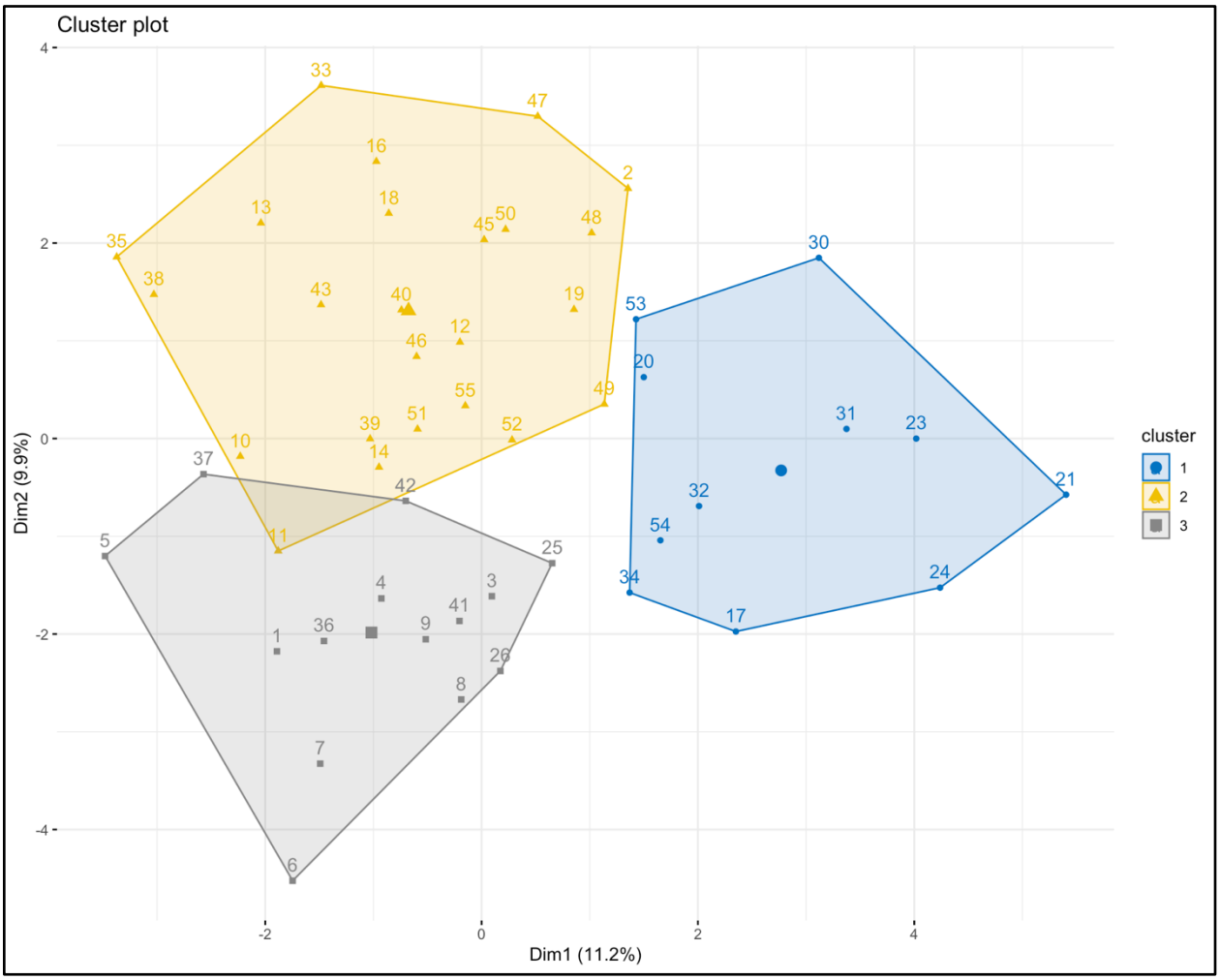


Figure 3.4

Three-cluster Solution for Possible Quality Verbal Feedback Typologies



CHAPTER 5

CONCLUSIONS

The conclusions and resulting implications of the three studies conducted and described in this dissertation are the result of years of research and data collection, in an attempt to examine quality verbal feedback in the field of music and pre-service teacher education. The purpose of Study 1 (see Chapter 2) was to create a tool which could be used to examine the psychological construct of quality verbal feedback by defining scale criteria based upon a thorough examination of the research literature surrounding formative assessment strategies and quality verbal feedback in the performance-based classroom. The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* was the resulting culmination of criteria carefully selected from the most common themes of quality verbal feedback found throughout the education research literature. The original scale consisted of 39 criteria across five domains. The research questions guiding the study included: (1) What are the psychometric qualities (i.e., validity, reliability, and precision) of the Pre-Service Music Teacher Verbal Feedback Evaluation Scale, (2) How do the verbal feedback criteria vary in difficulty, and (3) How does the rating scale category structure vary across each individual criterion? The data were analyzed using Rasch measurement, which is part of the family of Item Response Theory (IRT) and maintains the requirements of unidimensionality and invariant measurement (Engelhard, 2013). Specifically, the data were examined using the Many Facet Rasch Partial Credit Model (MFR-PC). Results showed a clear ordering of pre-service teachers, scale criteria, and domains based upon measures of difficulty. The *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* demonstrated

strong validity, reliability, and precisions measures based on data-to-model fit indices, the spread of logit scale locations, and separation statistics.

The purpose of Study 2 (see Chapter 3) was to examine the accuracy of pre-service music educators' self-assessment of the quality of their verbal feedback in the secondary-level instrumental ensemble setting using the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale.* After validation and re-evaluation of scale criteria, the revalidated scale to examine the quality of pre-service teachers' verbal feedback consisted of 35 items across the five domains. The research questions guiding the study included: (1) Overall, how accurate were pre-service music educators' perceptions of their verbal feedback when compared to content experts' perceptions, (2) How does accuracy vary across each item of the scale, and (3) How does accuracy vary across each domain of the scale? The Multifaceted Rasch Rater Accuracy Measurement Model (MFR-RA) was used to compare the self-assessment scores of pre-service music educators to the scores of a content expert in the field. Results showed that pre-service music educators' self-assessment scores had good model-to-data fit, and overall, pre-service teachers were able to accurately assess their performance in comparison to the assessment of a content expert. Therefore, the scale can be used as a reliable tool for self-assessment by pre-service music educators.

The purpose of Study 3 (see Chapter 4) was to examine if typologies of pre-service music educators exist based upon their ability to administer quality verbal feedback in the secondary-level instrumental music classroom. The research questions guiding the study included: (1) Does a meaningful verbal feedback typology exist based upon systematic differential item functioning (student-by-criterion) bias indices, (2) What are the predominant characteristics of the verbal feedback typologies, and (3) What can be concluded about the strengths and weaknesses in

teacher training programs, if any, from the patterns for typologies of pre-service teaching? A differential item functioning (DIF) analysis was performed to examine the bias indices of the latent variable, quality verbal feedback, being examined. A cluster analysis was also performed, and a three-cluster solution was chosen to represent pre-service music educators' typologies based on the criteria of quality verbal feedback. The three meaningful typologies of pre-service teachers resulting from the cluster analysis were: (a) learning objective and type focused, (b) context/audience, learning processed focused, and (c) clarity, goal, and time focused. Cluster 1, the learning objective and type focused pre-service teachers, accounted for 22.45% of the participant sample ($N = 11$). Cluster 2, the context/audience, learning processed focused teachers, accounted for 48.98% of the participant sample ($N = 24$). Cluster 3, the clarity, goal, and time focused teachers, accounted for 28.57% of the participant sample ($N = 14$). The substantive interpretation of the typologies represented by these three clusters show that certain criteria of quality verbal feedback tend to align with a specific type of pre-service music educator.

The ability for pre-service teachers to establish a repertoire of approaches to administer quality formative assessments through quality verbal feedback is often dependent on whether teacher educators prioritize the modeling of appropriate feedback pedagogy to their pre-service teachers (Brookhart, 2017). Perhaps pre-service teachers struggle to administer quality feedback because good pedagogical practices are not modeled by their professors during training, as "…mere exposure to subject matter alone does not ensure teacher effectiveness" (Howard & Aleman, 2008, p. 159). Perhaps music teacher training programs are too ambiguous with the expectations for their pre-service teachers during participation in authentic teaching experiences and contexts. Regardless of the reasons pre-service teachers struggle to transfer their content knowledge into practice, those charged with the task of teacher training need a way to assess

where the gaps in their programs are. It is not only important for pre-service teachers to understand the content that they are taught, but how to utilize various pedagogical practices in order to reach a diverse group of learners.

The research literature supports that measuring the growth of teacher effectiveness by gathering evidence-based documentation using high-quality, valid, and reliable assessment tools largely effects teacher quality and improvement, and therefore promotes more successful student outcomes (Flanders, 1965). These same types of assessments should also be utilized during pre-service teacher training. Not all pre-service teachers will have the same strengths and weaknesses when administering quality verbal feedback, but tools such as the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* allow teacher educators the opportunity to help their students locate and correct any gaps in their training. This feedback evaluation scale can be utilized in several ways, which may serve the purposes of different programs. First, the scale can be used in a pre-test/post-test format, where the teacher educator uses the measurement instrument to assess their students' teaching episodes. This would allow them to systematically examine their pre-service students at the beginning and end their course or semester. The data resulting could be used to determine growth and could indicate what weaknesses and misconceptions need to be further addressed in proceeding courses. Second, the scale can be used as a tool for self-reflection. Teacher educators often ask students to reflect on their practicum and lab teaching experiences, but do not often offer a systematic way for them to do so. This often results in pre-service teachers writing a reflection with no parameters as to what should be addressed or observed from their teaching episode. In order to perform the self-assessment successfully, pre-service teachers would need to video tape, watch their video, and complete the scale evaluation immediately after viewing. Recall of teaching performance, as discussed, has shown to yield

unreliable results, as memory recall is not always accurate. If pre-service teachers could perform a systematic self-assessment multiple times during a course, they would have the data collected from throughout the semester, giving them a more definitive diagnosis of their strengths and weaknesses. Last, the scale can be used as a measure of overall music teacher education program effectiveness. Schools of music could use the scale to gather data on their students' performances from entry into the program through their graduation. This data could be used to make necessary changes in course offerings and the structure of the curriculum in the music education department. It would also allow for more differentiation among pre-service teacher training, as not all students will have the same strengths or weaknesses.

In conclusion, the *Pre-Service Music Teacher Verbal Feedback Evaluation Scale* has been validated and revalidated, showing strong validity and reliability measures. The scale was constructed with the intention of using the criteria as a teaching tool for pre-service music educators studying to teach in the secondary-level instrumental music classroom. Moving forward, authors hope to continue the examination of differentiated instruction throughout music teacher training programs, using systematic methods of assessment, in order to develop more high-quality, well-rounded educators graduating from their pre-service training programs.

References

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education, 126*, 246–255. http://doi.org/10.2307/3345094

Alley, J. M. (1980). The effect of self-analysis of videotapes on selected competencies of music therapy majors. *Journal of Music Therapy*, 17, 113-132. https://doi.org/10.1093/jmt/17.3.113

Alton-Lee, A. (2003). *Quality teaching for diverse students in schooling: Best evidence synthesis*. Wellington, NZ.

Alton-Lee, A., & Nuthall, G. (1998). *Inclusive instructional design: Theoretical principles emerging from the understanding learning and teaching project.* Wellington, NZ.

Anderson, J. B., & Freiberg, H. J. (1995). Using self-assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly*, *22*(1), 77–91.

Asmus, E. (2000). The need for research in music education. *Journal of Music Teacher Education*, *10*(1), 5.

Ballantyne, J. (2007). Documenting praxis shock in early-career Australian music teachers: The impact of pre-service teacher education. *International Journal of Music Education, 25*(3), 181–191. http://doi.org/10.1177/0255761407083573

Bernabé-Valero, G., Blasco-Magraner, J. S., & Moret-Tatay, C. (2019). Testing motivational theories in music education: The role of effort and gratitude. *Frontiers in Behavioral Neuroscience*, *13*, 1–9. https://doi.org/10.3389/fnbeh.2019.00172

Bannister, N. A., & Linder, S. M. (2015). Recasting a traditionally summative assessment as an intentionally formative experience. *The Educational Forum*, *79*(2), 190–199. https://doi.org/10.1080/00131725.2014.1002594

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

Bergee, M.J. (1988). Use of an objectively constructed rating scale for the evaluation of brass juries: A criterion-related study. *Missouri Journal of Research in Music Education, 5*(5), 6-25.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463-482. https://doi.org/10.1016/S0883-0355(02)00004-6

Bernard, R. (2009). Uncovering pre-service music teachers' assumptions of teaching, learning, and music. *Music Education Research, 11*(1), 111–124. https://doi.org/10.1080/14613800802700974

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Book, C., Byers, J., & Freeman, D. (1983). Student expectations and teacher education traditions with which we can and cannot live. *Journal of Teacher Education*, *34*(1), 9–13. https://doi.org/10.1177%2F002248718303400103

Brand, M., & Burnsed, V. (1981). Music abilities and experiences as predictors of error-detection skill. *Journal of Research in Music Education*, *29*(2), 91–96. https://doi.org/10.2307/3345017

Brookhart, S. M. (2008). *How to give effective feedback to your students*. ASCD.

Brookhart, S. M. (2017). *How to give effective feedback to your students* (2nd ed.). ASCD.

Brookhart, S. M. (2017). Formative assessment in teacher education. In D. J. Clandinin & J. Husu (Eds.), *The SAGE Handbook of Research on Teacher Education* (pp. 927–943). https://doi.org/10.4135/9781526402042

Brophy, T. (2000) Assessing the developing child musician. GIA.

Burnett, P. C. (2002). Teacher praise and feedback and students' perceptions of the classroom environment. *Educational Psychology, 22*(1), 5–16. https://doi.org/10.1080/01443410120101215

Burrack, F., & Parkes, K. A. (2018). *Applying model cornerstone assessments in K-12 music: A research-supported approach*. Rowman & Littlefield.

Burrack, F., & Parkes, K. A. (2019). The development of standards-based assessments in music. In T. S. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education* (pp. 650–669). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190248093.013.28

Butler, A. (2001). Preservice music teachers' conceptions of teaching effectiveness, microteaching experiences, and teaching performance. *Journal of Research in Music Education*, *49*(3), 258–272. https://doi.org/10.2307%2F3345711

Chen, Y. H., Thompson, M. S., Kromrey, J. D., & Chang, G. H. (2011). Relations of student perceptions of teacher oral feedback with teacher expectancies and student self-concept. *Journal of Experimental Education, 79*(4), 452–477. https://doi.org/10.1080/00220973.2010.547888

Choy, D., Wong, A. F. L., Goh, K. C., & Ling Low, E. (2014). Practicum experience: Pre-service teachers' self-perception of their professional growth. *Innovations in Education*

and Teaching International, *51*(5), 472–482. https://doi.org/10.1080/14703297.2013.791552

Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education, 57*(1), 5–15. http://doi.org/10.1177/0022429409333405

Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.). (2008). *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed.). Routledge; Taylor & Francis Group; Association of Teacher Educators. https://doi.org/10.4324/9780203938690

Colwell, R. (1999). The 1997 assessment in music: Red flags in the sunset. *Arts Education Policy Review, 100*(6), 33–39. http://doi.org/10.1080/10632919909605996

Conroy, M. a, Sutherland, K. S., Snyder, A., Al-Hendawi, M., & Vo, A. (2009). Creating a positive classroom atmosphere: Teachers' use of effective praise and feedback. *Beyond Behavior*, 18–26.

Conway, C. (2002). Perceptions of beginning teachers, their mentors, and administrators regarding preservice music teacher preparation. *Journal of Research in Music Education*, *50*(1), 20–36. https://doi.org/10.2307/3345690

Conway, C. M. (2012). Ten years later: Teachers reflect on "Perceptions of beginning teachers, their mentors, and administrator regarding preservice music teacher preparation." *Journal of Research in Music Education*, *60*(3), 324–338. https://doi.org/10.1177/0022429412453601

Conway, C. M., & Hibbard, S. (2020). Pushing the boundaries from the inside. In Conway, C. M., Pellegrino, K., Stanley, A. M., & West, C. (Eds.), *The Oxford handbook of preservice music teacher education in the United States* (pp. 3-22). Oxford University Press.

Conway, C. M., Pellegrino, K., Stanley, A. M., & West, C. (Eds.). (2020). *The Oxford handbook of preservice music teacher education in the United States.* Oxford University Press.

Cranmore, J., & Wilhelm, R. (2017). Assessment and feedback practices of secondary music teachers: A descriptive case study. *Visions of Research in Music Education, 29*, 1–23. http://www-usr.rider.edu/~vrme/v29n1/visions/Cranmore.pdf

Crisp, B. R. (2007). Is it worth the effort? How feedback influences students' subsequent submission of assessable work. *Assessment & Evaluation in Higher Education, 32*(5), 571-581.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

Daniel, R. (2001). Self-assessment in performance. *British Journal of Music Education*, *13*(3), 215–226. https://doi.org/10.1017/S0265051701000316

Darling-Hammond, L. (1999) The case for university-based teacher education. In R. A. Roth (Ed.), *The role of the university in the preparation of teachers* (pp. 13-30). Falmer Press; Taylor & Francis Group.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, *13*(42), 1–48.

DeCarbo, N. (1984). The effect of years of teaching experience and major performance instrument on error detection scores of instrumental music teachers. *Ohio Music Education Association*, *11*, 28–32. https://www.jstor.org/stable/24127284

Deluca, C., & Klingerb, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy and Practice*, *17*(4), 419–438. https://doi.org/10.1080/0969594X.2010.516643

Dewey, J. E. (1910). *How we think.* D.C. Heath and Co., Publishers.

Dinkelman, T. (2003). Self-study in teacher education: A means and ends tool for promoting reflective teaching. *Journal of Teacher Education*, *54*(1), 6–18. https://doi.org/10.1177/0022487102238654

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). Cambridge University Press.

Duke, R. A., (2012). *Intelligent music teaching: Essays on the core principles of effective instruction.* Learning and Behavior Resources.

Duke, R. A., & Byo, J. L. (2011). *The habits of musicianship: A radical approach to beginning band*. University of Texas Center for Music Learning.

Duke, R. A., & Madsen, C. K. (1991). Proactive versus reactive teaching: Focusing observation on specific aspects of instruction. *Bulletin of the Council for Research in Music Education*, *108*, 1–14. http://www.jstor.org/stable/40318434

Edwards, A. S., Edwards, K. E., & Wesolowski, B. C. (2019). The psychometric evaluation of a wind band performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Research Studies in Music Education*, *41*(3), 343–367. https://doi.org/10.1177/1321103X18773103

Edwards, K. E., Edwards, A. S., & Wesolowski, B. C. (2018). Validation of a string performance rubric using the multifaceted Rasch measurement model. *Bulletin of the Council for Research in Music Education*, (215), 7–31.

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal, 1*(1), 3–18. https://doi.org/10.5070/l2.v1i1.9054

Engelhard, Jr., G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

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

Engelhard, Jr., G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Taylor & Francis.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*(4), 395–430.

Falsario, H. N., Muyong, R. F., & Nuevaespaña, J. S. (2014). Classroom climate and academic performance of education students. *DLSU Research Congress*, 1–7.

Feiman-Nemser, S. (2008). Teacher capacity for diverse learners: What do teachers need to know? In Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed., pp. 697-705). Routledge; Taylor & Francis Group; Association of Teacher Educators.

Flanders, N. A. (1965). *Teacher influence, pupil attitudes, and achievement*. U.S. Department of Health, Education, and Welfare, Office of Education.

Floden, R. E. (2008). Improving methods for research on teacher education. In Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed., pp. 1183-1188). Routledge; Taylor & Francis Group; Association of Teacher Educators.

Freiberg, H. J. (1987). Teacher self-evaluation and principal supervision. *National Association of Secondary School Principals*, *71*(498), 85–92. https://doi.org/https://doi.org/10.1177/019263658707149814

Freiberg, H. J., Waxman, H. C., & Houston, W. R. (1987). Enriching feedback to student-teachers through small group discussion. *Teacher Education Quarterly*, *14*(3), 71–82.

Gelfuso, A., & Dennis, D. V. (2014). Getting reflection off the page: The challenges of developing support structures for pre-service teacher reflection. *Teaching and Teacher Education*, *38*, 1–11. https://doi.org/10.1016/j.tate.2013.10.012

Gickling, E. E., & Thompson, V. P. (1985). A personal view of curriculum-based assessment. *Exceptional Children*, *52*(3), 205–218. https://doi.org/https://doi.org/10.1177%2F001440298505200302

Gonzo, C., & Forsythe, J. (1976). Developing and using videotapes to teach rehearsal techniques and principles. *Journal of Research in Music Education*, *24*(1), 32–41. https://doi.org/10.2307/3345064

Hale, C. L., & Green, S. K. (2009). Six key principles for music assessment. *Music Educators Journal*, *95*(4), 27–31. http://doi.org/10.1177/0027432109334772

Haning, M. (2021). Identity formation in music teacher education: The role of the curriculum. *International Journal of Music Education*, *39*(1), 39–49. https://doi.org/10.1177/0255761420952215

Hash, P. M., (2020). A historical overview of music teacher education in the United States. In Conway, C. M., Pellegrino, K., Stanley, A. M., & West, C. (Eds.), *The Oxford handbook of preservice music teacher education in the United States* (pp. 45-66). Oxford University Press.

Hattie, J. (2012). *Visible learning for teachers*. Routledge.

Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback*. Routledge.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, *24*(1), 393–446. https://doi.org/10.3102/0091732x024001393

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hounsell, D. (2003). Student feedback, learning and development. In Slowey, M. & Watson, D. (Eds.), *Higher education and the lifecourse* (pp. 67-78). McGraw-Hill Education.

Houston, W. R. (2008). Settings are more than sites. In Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed., pp. 388-393). Routledge; Taylor & Francis Group; Association of Teacher Educators.

Howard, T. C., & Aleman, G. R. (2008). Teacher capacity for diverse learners: What do teachers need to know? In Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed., pp. 157-174). Routledge; Taylor & Francis Group; Association of Teacher Educators.

Iofee, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing interval covariate shift. *Proceedings of the 32nd International Conference on Machine Learning.* Retrieved from https://arxiv.org/abs/1502.03167

Jones, H. (1986). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school solo vocal performance (Doctoral dissertation, University of Oklahoma). *Dissertation Abstracts International*, *47*, 1230A.

Jordan-DeCarbo, J. (1986). A sound-to-symbol approach to learning music. *Music Educators Journal*, *72*(6), 38–41. https://doi.org/https://doi.org/10.2307%2F3399067

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–154). Lawrence Erlbaum Associates Publishers.

Kaschub, M., & Smith, J. (2014). Music teacher education in transition. In Michele Kaschub & J. Smith (Eds.), *Promising practices in 21st century music teacher education* (pp. 3–24). Oxford University Press.

Keep, B. (2019). *Survey says…: How to write surveys to get the information you want*. https://www.benjaminkeep.com/product-page/survey-says-how-to-write-surveys-to-get-the-information-you-want

Kimpton, P., & Kimpton, A. (2019). Making assessment meaningful, measurable, and manageable in the secondary music classroom. In T. S. Brophy (Ed.), *Oxford handbook of assessment policy and practice in music education* (pp. 324–350). https://doi.org/10.1093/oxfordhb/9780190248130.013.52

Legette, R. M., & McCord, D. (2015). Pre-service music teachers perceptions of teaching and teacher training. *Contributions to Music Education, 40*, 163–176.

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. 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*(1), 85–106.

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

MacLeod, R. B., & Nápoles, J. (2013). Preservice teachers' perceptions of teaching effectiveness during teaching episodes with positive and negative feedback. *Journal of Music Teacher Education, 22*(1), 91–102. https://doi.org/10.1177%2F1057083711429851

Masters, G. N. (1982). Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272

Madsen, C. K., Standley, J. M., & Cassidy, J. W. (1989). Demonstration and recognition of high and low contrasts in teacher intensity. *Journal of Research in Music Education*, *37*(2), 85-92. https://doi.org/10.2307%2F3344700

Morrison, G. R., Ross, S. J., Morrison, J. R., & Kalman, H. K. (2019). *Designing effective instruction* (8th ed.). John Wiley & Sons.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: perceptions of students and academics. *Assessment and Evaluation in Higher Education*, *42*(2), 266–288. https://doi.org/10.1080/02602938.2015.1103365

Musselwhite, D. J., & Wesolowski, B. C. (2018). 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. *Journal of Research in Music Education*, *66*(3), 338–358. https://doi.org/10.1177/0022429418793645

Musselwhite, D. J., & Wesolowski, B. C. (2021). Evaluating the psychometric qualities of the edTPA in the context of pre-service music teachers. *Research Studies in Music Education*, *43*(1), 39–58. https://doi.org/10.1177/1321103X19872232

Natriello, G. (1987). The impact of evaluation process on students. *Educational Psychologist*, *22*(2), 155–175.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Odendaal, A. (2013). *Perceptual learning style as an influence on the practising of instrument students in higher music education*. Cape Town, South Africa.

Odendaal, A. (2016). (Mis)matching perceptual learning styles and practicing behavior in tertiary level Western Classical instrumentalists. *Psychology of Music*, *44*(3), 353–368. https://doi.org/10.1177/0305735614567933

Orrell, J. (2006). Feedback on learning achievement: Rhetoric and reality. *Teaching in Higher Education, 11*(4), 441–456. https://doi.org/10.1080/13562510600874235

Parkes, K. A. (2020). Student teaching and certification assessments. In Conway, C. M., Pellegrino, K., Stanley, A. M., & West, C. (Eds.), *The Oxford handbook of preservice music teacher education in the United States* (pp. 231-252). Oxford University Press.

Parkes, K. A., Ritcher, G. K., & Doerksen, P. F. (2019). Measuring dispositions in preservice music educators In T. S. Brophy (Ed.), *The Oxford Handbook of Assessment Policy and Practice in Music Education* (Vol. 1, pp.869-899). Oxford University Press.

Perpignan, H. (2003). Exploring the written feedback dialogue: a research, learning and teaching practice. *Language Teaching Research, 7*(2), 259–278. https://doi.org/10.1191%2F1362168803lr125oa

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Rasch, G. (1960/1980). *Probabilistic models for some intelligent an attainment tests.* MESA Press.

Richards, C., & Killen, R. (1993). Problems of beginning teachers: Perceptions of pre-service music teachers. *Research Studies in Music Education, 1*(1), 40–51. http://doi.org/10.1177/1321103X9300100105

Romesburg, H. C. (1984). *Cluster analysis for researchers.* Lifetime Learning.

Rosaen, C., & Florio-Ruane, S. (2008). The metaphors by which we teach: Experience, metaphor, and culture in teacher education. In Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D. J., & Demers, K. E. (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed., pp. 706-731). Routledge; Taylor & Francis Group; Association of Teacher Educators.

Rosenthal, R. K. (1985). Improving teacher effectiveness through self-assessment: A case study. *Update*, *3*, 17–21.

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

Russell, V. (2009). Corrective feedback, over a decade of research since Lyster and Ranta (1997): Where do we stand today? *Electronic Journal of Foreign Language Teaching, 6*(1), 21–31.

Rutkowski, J., & Miller, M. S. (2003). The effect of teacher feedback and modeling on first graders' use of singing voice and developmental music aptitude. *Bulletin of the Council for Research in Music Education, 156*, 1–10.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *The Journal of Higher Education, 54*(1), 60–79. https://doi.org/10.1080/00221546.1983.11778152

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 77–84. https://doi.org/10.1080/0969595980050104

Saliba, K., & Barrett, J. R. (1993). Undergraduate methodology specific or general. *Journal of Music Teacher Education*, *3*(1), 23–28. https://doi.org/10.1177/105708379300300105

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education*, *61*(4), 339–349. https://doi.org/10.1177%2F0022487110369555

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 45–59. https://doi.org/10.18148/srm/2010.v4i1.2682

Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education, 45*(2), 259–272. http://doi.org/10.2307/3345585

Senko, C., & Harackiewicz, J. M. (2005). Regulation of achievement goals: The role of competence feedback. *Journal of Educational Psychology, 97*(3), 320–336. https://psycnet.apa.org/doi/10.1037/0022-0663.97.3.320

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polygamous data. *BMC Medical Research Methodology, 8*, 1-11. https://doi.org/10.1186/1471-2288-8-33

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

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education, 57*(3), 217–235. http://doi.org/10.1177/0022429409343549

Tindale, R. S., Kulik, C. T., & Scott, L. A. (1991). Individual and group feedback and performance: An attributional perspective. *Basic and Applied Social Psychology, 12*(1), 41–62. https://doi.org/10.1207/s15324834basp1201_4

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners*. Association for Supervision and Curriculum Development.

Tomlinson, C. A. (2014). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). ASCD.

Walker, R. J. (2008). Twelve characteristics of an effective teacher: A longitudinal, qualitative, quasi-research study of in-service and pre-service teachers' opinions. *Educational Horizons, 87*(1), 61–68.

Ward, J. H., (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Wesolowski, B. C. (2015). Tracking student achievement in music performance. *Music Educators Journal*, *102*(1), 39–47. https://doi.org/10.1177/0027432115589352

Wesolowski, B. C. (2016). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music*, *44*(3), 324–339. https://doi.org/10.1177/0305735614567700

Wesolowski, B. C. (2017). Exploring rater cognition: A typology of raters in the context of music performance assessment. *Psychology of Music*, *45*(3), 375–399. https://doi.org/10.1177/0305735616665004

Wesolowski, B. C. (2019). Item response theory and music testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education* (pp. 479–503). Oxford University Press.

Wesolowski, B. C. (2020). Validity, reliability, and fairness in classroom tests. In K. A. Parkes & F. Burrack (Eds.), *Developing and applying assessments in the music classroom* (1st ed., pp. 82–102). https://doi.org/10.4324/9780429202308-5

Wesolowski, B. C., Athanas, M. I., Burton, J. S., Edwards, A. S., Edwards, K. E., Goins, Q. R., Irby, A. H., Johns, P. M., Musselwhite, D. J., Parido, B. T., Sorrell, G. W., & Thompson, J. E. (2018). Judgmental standard setting: The development of objective content and performance standards for secondary-level solo instrumental music assessment. *Journal of Research in Music Education*, *66*(2), 224–245. https://doi.org/10.1177/0022429418765482

Wesolowski, B. C., & Ng, A. (2020). *The development of a predictive opportunity-to-learn model using machine learning*. Reston, VA.

Wesolowski, B. C., & Wind, S. A. (2019). *Examining parents' perceptions of self-regulatory and self-determinative screen-based learning behaviors for adolescents participating in private music study*. Los Angeles, CA.

Wesolowski, B. C., Wind, S. A., & Engelhard, Jr., G. (2016). Examining rater precision in music assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception*, *33*(5), 662–678. https://doi.org/10.1525/MP.2016.33.5.662

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wind, S. A., Ooi, P. S., & Engelhard, Jr., G. (2018). Exploring decision consistency and decision accuracy across rating designs in rater-mediated music performance assessments. *Musicae Scientiae*, 1–21. https://doi.org/10.1177%2F1029864918761184

Wind, S. A., & Wesolowski, B. C. (2018). Evaluating differential rater accuracy over time in solo music performance assessment. *Bulletin of the Council for Research in Music Education*, (215), 33–55.

Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, *22*, 34–39. http://doi.org/10.1177/025576149302200106

White, S. (2007). Investigating effective feedback practices for pre-service teacher education students on practicum. *Teaching Education, 18*(4), 299–311. https://doi.org/10.1080/10476210701687591

Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, *16*(2), 153—160.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

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

Wright, B. D., & Stone, M. H. (1999). *Measurement Essentials* (2nd ed.). Wide Range.

Wright, B. D., & Tennant, A, (1996). Sample size again. *Rasch Measurement Transactions, 9*(4), 468.

Yarbrough, C. (1987). The relationship of behavioral self-assessment to the achievement of basic conducting skills. *Journal of Research in Music Education*, *35*(3), 183–189. https://doi.org/10.2307/3344960

Yarbrough, C., Wapnick, J., & Kelly, R. (1979). Effect of videotape feedback techniques on performance, verbalization, and attitude of beginning conductors. *Journal of Research in Music Education*, *27*(2), 103–112. https://doi.org/10.2307/3344896

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education, 50*, 245–255. http://doi.org/10.2307/3345801

Zhukov, K. (2007). Student learning styles in advanced instrumental music lessons. *Music Education Research*, *9*(1), 111–127. https://doi.org/10.1080/14613800601127585